UNIVERSITÉ DE NANTES

___

FACULTÉ DE MÉDECINE

___

Année: 2020 N°

THÈSE

pour le

DIPLÔME D'ÉTAT DE DOCTEUR EN MÉDECINE

DES de Biologie Médicale

par

Martin Broly

Né le 31 Août 1990 à Lille (59)

___

Présentée et soutenue publiquement le 30 Novembre 2020

___

Functional genomics in a diagnostic setting: ANEVA-DOT pipeline refinement and application to the Congenital Heart Disease GEnetic NEtwork Study cohort

___

Président du jury : Monsieur le Professeur Stéphane Bézieau

Directrice de thèse : Madame le Professeur Tuuli Lappalainen

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Tuuli Lappalainen for giving me the opportunity to carry out this project and for providing invaluable guidance and support all along. I could not have imagined having a better advisor and mentor. Thank you for your patience, kindness, enthusiasm, immense knowledge and vision for your field, academia and society that have deeply inspired me.

I am extending my thanks to everyone from the Lappalainen lab, especially Jonah, Marcello, Dafni, Silva, Paul, Vardiella, Molly, John, Elise, Anu, Kristina and Kathryn for their encouragement, tips, genuine help and kindness.

A Monsieur le Professeur Stéphane Bézieau qui me fait l'honneur de présider ce jury. Merci pour votre soutien dans la construction de ce projet et d'autres, et plus largement pour vos conseils et votre bienveillance. Recevez ici l'expression de toute ma reconnaissance.

A Monsieur le Professeur Jean-Marie Bard, qui me fait l'honneur de juger ce travail. Veuillez trouver ici l'expression de mes sincères remerciements.

A Monsieur le Professeur Thierry Le Tourneau, que je remercie d'avoir accepté de faire partie de mon jury de thèse.

A mes parents, merci pour votre soutien à toute épreuve tout au long de mon cursus et depuis toujours. Merci pour vos sacrifices et pour m'avoir inspiré sur le plan personnel et professionnel. Je ne serais ni dans cette spécialité, ni arrivé si loin, ni qui je suis sans vous.

A mes soeurs, Joséphine, Charlotte et Elyette, merci d'avoir toujours confiance en moi quoi qu'il arrive (et de toujours trouver que ce que je fais est génial !). Je suis fier de notre fratrie si soudée.

A mes grands-parents, merci d'être des modèles pour moi et nous tous, vos petits-enfants.

A mes amis de toujours, Simon, Nicolas, Thomas et Marie. Merci pour votre amitié indéfectible malgré la distance, tous nos souvenirs communs et tous ceux à venir.

A Paul-Henri, pour ta disponibilité permanente à refaire le monde ensemble

A Marc, ami de la première heure, pour toutes ces années de fac, nos voyages, nos projets et tout ce que nous avons vécu ensemble

A Antoine, Margaux, Alexia, Zélie, Guillaume, Morgane ; la coloc Saint Jacques, pour cette première année d'internat et les suivantes !

A Justine, Antoine, Vincent, Louise, Louise, Claire, Adeline, Anaïs et tous ceux qui ont égayé ces années d'internat

Table of Contents

Table of figures

List of abbreviations

AARSD1: Alanyl-TRNA Synthetase Domain Containing 1

ACMG: American College of Medical Genetics and Genomics

AC: Allelic Count

AE: Allelic Expression

aeSNV: Allelic expression Single Nucleotide Variant

AF: Allele Frequency

ASE: Allele-Specific Expression

ASPN: Asporin

AMP: Association for Molecular Pathology

ANEVA: ANalysis of Expression Variation

ANEVA-DOT: ANalysis of Expression Variation-Dosage Outlier Test

ARTAORT: Aorta tissue type

BAM: Binary Alignment Map

BMP: Bone morphogenetic proteins

BWA-MEM: Burrows-Wheeler Aligner

CADD: Combined Annotation Dependent Depletion

CHD: Congenital Heart Defects

CHD GENES: Congenital Heart Disease GEnetic NEtwork Study

CI: Confidence Interval

ELN: Elastin

ENG: Endoglin

FDR: False Discovery Rate

GATK: Genome Analysis ToolKit

GTEx: Genotype-Tissue Expression

HRTAA: Atrial Appendage tissue type

HRTLV: Left Ventricle tissue type

HTS: High-Throughput Sequencing

MAP: Mitogen-Activated Protein

MAP2K1: Mitogen-Activated Protein Kinase Kinase 1

MNP: MultiNucleotide Polymorphism

NGS: Next-Generation Sequencing

o/e ratio: Observed over Expected number of loss-of-function variants ratio

PCGC: Pediatric Cardiac Genomics Consortium

PGM1: Phosphoglucomutase 1

phASER: phasing and Allele Specific Expression from RNA-seq

PTGES3L: Prostaglandin E synthase 3 (cytosolic)-like

RBFOX2: RNA Binding Fox-1 Homolog 2

RNA-seq: RNA-sequencing

RIN: RNA integrity number

ROI: Region Of Interest

SNV: Single Nucleotide Variant

SMAD6: SMAD Family Member 6

TGF-β: Transforming growth factor beta

VCF: Variant Call Format

VEP: Variant Effect Predictor

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

1. Introduction

1.1 The high-throughput era

High-throughput Sequencing (HTS) technologies provide simultaneous testing of multiple genes through the generation of millions of DNA sequences in hours, at great depth and decreasing costs (1). The fast success of NGS in research revolutionized the field of genomics and its implementation in patient's care routine practice already transformed the field of medical diagnosis (2,3). High resolution, low biases and detection power make possible discoveries unachievable with previous technologies and now serve many purposes (4). Applications of NGS technologies in a clinical setting range from custom panels (multi-gene), whole-exome sequencing (WES), whole-genome sequencing (WGS), RNA-sequencing (RNA-seq) and epigenetic testing (5).

Single-gene testing or custom panels are indicated when the clinical features for a patient are typical of a particular disorder and the association between the disorder and a few specific genes is well established (6). However, cardiomyopathies, epilepsy, congenital muscular dystrophy or X-linked intellectual disability (7) for instance are more complex diseases, mainly because of clinical variability and genetic locus heterogeneity, with many potential causal genes and variants. Such diseases now require large gene panels or WES to give clinicians a chance to make a diagnosis because other methods would be costly and time consuming. For inherited cancer risk evaluation, multi-gene panels may include high-penetrance genes as well as associated genes with a moderate increase in risk (7) but targeted exome sequencing is becoming increasingly popular for assessing the full diversity of cancer-related genes and the identification of subclonal mutations (8,9). The major benefit of WGS is the almost complete coverage of the genome, including structural variants, promoters and regulatory regions with a uniform coverage in the regions of interest (ROIs) but at the cost of having a lower depth of sequencing. Still, the chances of not detecting disease-causing variants due to technical errors is much lower with WGS than WES (10–12). RNA-seq consists of an in-depth RNA analysis through NGS technologies and has become the go-to technology for whole transcriptome data generation (13). It enables

detection of novel gene isoforms, gene fusions, splice and structural variants, gene expression quantification and allele-specific expression quantification (14).

1.2 Diagnostic and patient's care improvements

The usefulness of performing NGS in a clinical setting varies depending on the disorder. First, once potential causative variant(s) have been identified, useful information for variant interpretation can be gathered from existing databases such as gnomAD for population data at the variant and gene level (15) but also phenotypic and disease databases such as OMIM (16) and Orphanet (17) for gene specific prognostication (18,19). Second, if a diagnosis is achieved, it allows an end to an expensive and stressful diagnostic odyssey. Reaching a molecular diagnosis relieves the guilt that parents may face in the absence of a firm diagnosis (20,21) and helps them accept their child's condition. Third, it allows for anticipatory management of other comorbidities that an individual may be susceptible to by assessing other organs for possible complications (22). More importantly, it facilitates genetic counseling and allows for more accurate estimates of recurrence risk in the family because identification of the molecular etiology can be used to guide subsequent pregnancies, either through prenatal diagnostics or preimplantation genetic diagnosis (23,24). In some situations, it also allows for identification of other at-risk family members and any available treatment can be instituted in the presymptomatic phase (21). For instance, identification of a mutation in a gene causing long QT syndrome in the proband can allow identification of other at-risk family members, who can then be monitored to implement appropriate care (25,26). Lastly, it is possible that targeted molecular therapy may be available for a specific genetic mutation that helps to improve the patient's symptoms. For example, some individuals with vascular malformations involving somatic mutations in the AKT3-PI3K-mTOR pathway have been successfully treated with mTOR-inhibitors such as rapamycin (27,28).

1.3 Congenital Heart Diseases

Congenital heart disease (CHD) defines a large set of structural and functional deficits that arise during cardiac embryogenesis such as coarctation of the aorta, aortic or pulmonary stenosis, atrial / ventricular / atrioventricular septal defect, tetralogy of Fallot , hypoplastic left heart, mitral atresia, patent ductus arteriosus, persistent truncus arteriosus, transposition of the great arteries… (29). It is the most common type of birth defect, occuring in ~1% of newborns and accounting for one third of all major congenital anomalies (30). Although environmental exposures are relevant (29,31), epidemiologic studies strongly suggest genetic factors as a cause of CHD. Oyen *et al.* showed that after exclusion of chromosomal defects, the population-associated risk given a positive family history of CHD was 4.2% and that parental consanguinity is associated with a 2- to 3-fold increased offspring risk of CHD (32). It is likely that there are a large number of causative genes overall, and the expected frequency of causative variants for any one gene is low (33). Genes known to cause CHD already include transcription factors (34–36), signalling molecules (29,37,38) and chromatin modifiers (39–41) which direct the temporal and spatial expression of genes during cardiac development but the molecular basis underlying congenital heart defects largely remain to be elucidated.

1.4 Challenges remain

The data generated with HTS technologies have grown spectacularly and raw data volumes typically range between a few gigabytes per sample for WES and transcriptome data to around 100 gigabytes for WGS (42). Close to 100,000 SNVs and indels per individual are detected on WES (43) and millions on WGS (44). The main hurdle that remains is now to translate this massive amount of genomic data into genomic medicine. Connecting variants to disease and distinguishing disease-causing genetic variants from candidate variants still is a complex task (45,46). Even though early steps are highly automated, the final aspects are not and rely on expert review and human interpretation instead. In a joint effort, The American College of Medical Genetics and Genomics (ACMG) and the Association for

Molecular Pathology (AMP) have described guidelines for classifying variants into five categories based on criteria using typical types of variant evidence (e.g., population data, computational data, functional data, segregation data) (47). Also, many computational approaches have been developed to tackle the variant prioritization problem regardless or not of the wider context of gene prioritization (48).

RNA-sequencing based methods provide diverse measurements to assess the functional effects of genetic variants and have the potential for application across diverse areas of human health, including disease diagnosis, prognosis and therapeutic selection (14). For instance, splicing alterations or structural variants can give rise to alternative transcript variants that are implicated in human diseases, including developmental disorders (49), neurodegenerative disorders (50,51), neuromuscular disorders (52) and cancers (53). One recent approach to improve the interpretation of genetic variation and gene prioritization is to integrate functional genomic information in the form of allele-specific expression (ASE) obtained from disease-relevant tissues with genomic data (54–57). ASE data can be used to quantify expression variation between two haplotypes of a diploid individual distinguished by heterozygous sites (54). In 2019, Mohammadi *et al.* introduced the ANalysis of Expression Variation (ANEVA) to quantify genetic variation in gene dosage from allelic expression data in a population (58) and developed a framework to use these variance estimates in a dosage outlier test (ANEVA-DOT) that can be incorporated in rare disease NGS pipelines to use RNA-sequencing data more effectively. Applying this gene-prioritization method to ASE data from 70 Mendelian muscular disease patients showed accuracy in detecting genes with pathogenic variants in previously resolved cases and led to one confirmed and several potential new diagnoses (58).

1.5 Study questions

The purpose of this work was first to refine the general ANEVA-DOT pipeline to improve its accuracy by using phASER, a software providing measures of haplotypic expression that increases power and accuracy in studies of allelic expression (59). Second, we apply this

novel framework to a cohort of patients with Congenital Heart Diseases (CHDs) in order to identify variants responsible for patients' diseases.

2. Material and methods

2.1 The Congenital Heart Disease GEnetic NEtwork Study

The Pediatric Cardiac Genomics Consortium (PCGC) designed the Congenital Heart Disease GEnetic NEtwork Study (CHD GENES) to provide phenotype and genotype data for a large congenital heart defects (CHDs) cohort (60). The PCGC cohort probands were recruited from five main sites (Children's Hospital of Philadelphia, Columbia University Medical Center, Harvard Medical School including Boston Children's Hospital and Brigham and Women's Hospital, Icahn School of Medicine at Mount Sinai, and Yale School of Medicine) and four satellite sites (Children's Hospital of Los Angeles, Cohen Children's Medical Center, University College London, and University of Rochester Medical Center) in the United States and the United Kingdom (CHD Genes: NCT01196182) from November 2010 onwards. Recruitment methods were center-specific, but cardiac diagnoses were confirmed by review of imaging (e.g., echocardiogram) and operative reports. CHD diagnoses were assigned using the International Paediatric and Congenital Cardiac Codes (http://www.ipccc.net/) and manually reviewed by the main investigators. In addition, information on cases and their parents was obtained during subject and family interviews (60). Information on genetic testing, genetic physical exams, and extracardiac malformations was abstracted from medical records. Patient's recruitment, CHD diagnoses and information retrieval from medical records was performed by the PCGC consortium.

2.2 Study population

Our cohort is a subset of the CHD GENES cohort for which heart tissue samples were available. 397 heart tissue samples belonging to 350 patients were obtained from surgery-related discarded tissue. 257 samples corresponding to 220 patients remained after excluding RNA-seq samples that did not have a corresponding VCF file and excluding patients having more than 10% missing genotype calls. Samples originated from various anatomical regions of the heart but were classified into one of three categories according to their location: Atrial Appendage (HRTAA), Aorta (ARTAORT) and Left Ventricle (HRTLV) to

follow the Genotype-Tissue Expression (GTEx) (61) tissue collection sites. Patients were mainly newborn with a median age at consent of 6 months old (range: 0-21 years old). 64 samples were Aorta (ARTAORT) samples, 100 were Atrial Appendage (HRTAA) samples and 94 were Left Ventricle (HRTLV) samples (Figure 1).

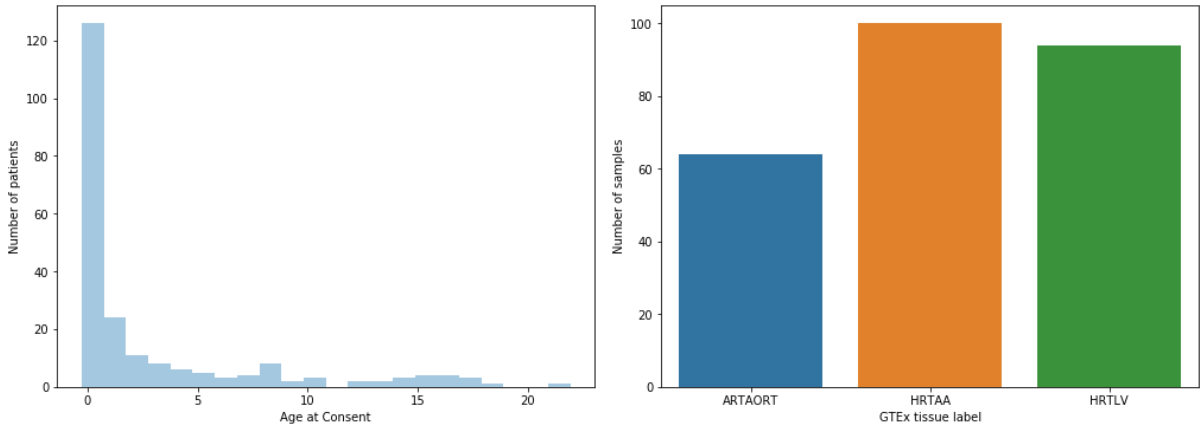


Figure 1: Age at consent and samples tissue of origin.

Left: Age at consent distribution for our patients. Right: Number of samples according to the tissue of origin. ARTAORT: Aorta, HRTAA: Atrial Appendage, HRTLV: Left Ventricle of the heart.

2.3 Dosage outlier testing from allelic expression

ANEVA is used to quantify genetic variation in gene dosage from allelic expression data in a population. It allows biologically interpretable estimates of genetic variation in gene expression within a population to be derived from ASE read count data (58). More specifically, for each gene, it estimates V^G, the expected variance in the dosage that is due to interindividual genetic differences within a population (58). Mohammadi *et al.* also developed the ANEVA-Dosage Outlier Test (ANEVA-DOT) to identify genes with an unusually strong effect on gene dosage using V^G scores and the observed gene expression pattern derived from ASE within one individual's tissue (58). If the observed allelic imbalance in an individual is not consistent with dosage variation in the general population, the gene is more likely affected by a heterozygous genetic variant. ANEVA-DOT is available as an R package and was performed for each sample independently. Only sites where the total coverage (REF count + ALT count) was > 10 were considered for testing. V^G scores based

on heart tissue data from GTEx version 8, available from https://github.com/PejLab/ANEVA-DOT_reference_datasets/tree/master/, were used. This provided a list of significant ANEVA-DOT outlier genes at 5% FDR for each sample (hereafter outlier genes).

2.4 RNA-seq data generation

RNA was purified from RNA later-treated frozen samples, using Trizol (Life Technologies). RNA (RIN > 5) was converted into cDNA and into RNA-seq libraries as described before (62). In brief, purified poly-A RNA that had gone through two rounds of oligo-dT selection was converted into cDNA and then made into RNA-seq libraries. Libraries were sequenced (Illumina HiSeq 2000 or Illumina HiSeq 2500, 50-base paired-end reads) to a target depth of > 20 million reads (median, 57 million reads). Mitochondrial reads were discarded using Samtools (63). Reads were aligned using TopHat 1.4 (64) to the Hg19 build of the human genome. RNA-seq data generation and alignment with TopHat was performed by the PCGC consortium.

Because ANEVA V^G scores were calculated based on data from GTEx version 8, thus depending on the Hg38 build of the human genome, we re-aligned RNA-seq data to Hg38. Briefly, RNA-seq BAM files were converted back to unaligned BAM files using Picard RevertSam version 2.21.3 (65) and subsequently to FASTQ files using bedtools v2.25.0 (66) bamtofastq utility before being realigned to the Hg38 build of the human genome using STAR 2.7.3 (67) in two-pass mode and the --waspOutputMode parameter set to “SAMtag”. This parameter is a re-implementation of the original WASP algorithm developed by Bryce van de Geijn *et al.* (68). Briefly, when aligning reads that contain the non-reference allele to a reference genome, these may not map uniquely or might map to a different (incorrect) location in the genome (69). WASP overcomes mapping bias by identifying mapped reads that overlap SNVs. For each read that overlaps a SNV, the allele that is present in the read is changed to match the SNV’s other allele, and the read is remapped. If a remapped read does map to exactly the same location, STAR will add the vW tag to the SAM output: vW:i:1 meaning alignment passed WASP filtering (70). If not, the read is tagged with vW:i:[2-7]

depending on the WASP quality criteria that failed. Mapping bias correction using WASP is the gold standard procedure for ASE analyses (71).

2.5 WES and WGS data generation

Exomes were captured and sequenced at the Yale Center for Genome Analysis, as previously described (39). In brief, gDNA isolated from venous blood was captured with the NimbleGen v2.0 exome capture reagent (Roche) and sequenced (Illumina HiSeq 2000, 75 base paired-end reads) to a mean read depth of 107 (72). Reads were aligned to the Hg19 build of the human genome using Novoalign (Novocraft), and further processed using the Genome Analysis ToolKit (GATK) Best Practices workflows (34–36). Single nucleotide variants and indels were called with GATK HaplotypeCaller (73).

Whole-genome sequencing was performed at the Baylor College of Medicine Genomic and RNA Profiling Core, the New York Genome Center (NYGC) Genomic Research Services, and the Broad Institute for Genomic Services as previously described (74). Briefly, gDNA from venous blood or saliva was prepared for sequencing using a PCR-free library preparation. All samples were sequenced on an Illumina Hi-Seq X Ten system with 150-bp paired reads to a median depth of > 30× per individual. Reads were aligned to the Hg19 or Hg38 build of the human genome with the Burrows-Wheeler Aligner (BWA-MEM) (75). GATK best practices recommendations were implemented for base quality score recalibration, indel realignment, and duplicate removal (73).

Patients VCF files were reviewed by a team of experts from New York-Presbyterian Hospital/Columbia University Medical Center Clinical Genetics and Genomics department. 213/220 patients did not have a genetic diagnosis after analysis of their WES or WGS data (60). WES / WGS data generation and VCF files review were performed by the PCGC consortium.

2.6 WES and WGS preprocessing

VCF files were lifted over from the Hg19 to Hg38 build of the human genome using Crossmap v0.4.2 (76). 70 patients were sequenced twice, once as WES and a second time

as WGS. For these patients, both files were concatenated into a unique file, dropping duplicate records. Then, the number of missing genotypes calls per patient was assessed and patient's sex was empirically confirmed as part of our quality control check using Plink v2.00 (77), an open-source whole genome association analysis toolset. Patients having more than 10% missing genotype calls were removed from downstream analysis.

Two VCFs file versions were then made for each patient depending on the subsequent destination pipeline for ASE data generation:

- ASEReadCounter pipeline: MultiNucleotide Polymorphisms (MNP) were split into Single Nucleotide Polymorphisms using GATK VariantsToAllelicPrimitives tool and only SNVs were selected using GATK SelectVariants tool (73).
- phASER pipeline: VCF files were phased for each patient using the 1000 genomes dataset (78) as a reference panel +/- parents information when it was available using SHAPEIT4, a fast and accurate method for phasing high coverage sequencing data (79).

2.7 Allele Specific Expression data generation

Using RNA-seq data, the general idea of generating ASE data corresponds to the tallying of the number of reads that carry the reference allele and the number of reads that support the alternate allele, denominated as REF count and ALT count, respectively. We tested for any identity mismatches between VCF files and RNA-seq data using QTLtools v1.2 MBV utility (80) and corrected them where appropriate.

ASE was generated in two different ways as part of the two pipelines we wanted to evaluate:

- **ASEReadCounter_wasp pipeline:** Allele-specific counts were generated from uniquely mapping reads using the GATK ASEReadCounter (73) utility taking a patient's VCF file (pre-processed as described above "2.6 WES and WGS preprocessing" > "ASEReadCounter pipeline) and the corresponding RNA-seq BAM

file filtered to keep only reads that passed WASP quality criteria (tagged with vW:i:1) as input. The output ASE data was subsequently annotated using gencode v26 (81).

- **Phaser_wasp pipeline:** Allele-specific counts were generated from uniquely mapping reads using phASER (59). phASER addresses the limitation of double counting of reads with tools like ASEReadCounter if multiple heterozygous variants in the same gene are covered by the same read (59). Instead, phASER integrates read counts across phased variants in order to quantify the expression of phased haplotypes by reporting the number of unique reads that map to each, which can greatly improve the power to detect ASE (59). For any given sample, phASER takes a VCF file (pre-processed as described above “2.6 WES and WGS preprocessing” > “phASER pipeline) and the corresponding RNA-seq BAM file filtered to keep only reads that passed WASP quality criteria (tagged with vW:i:1) and generates haplotype level ASE data. The output ASE data was subsequently annotated using gencode v26 (81).

2.8 Putative causal variant screening

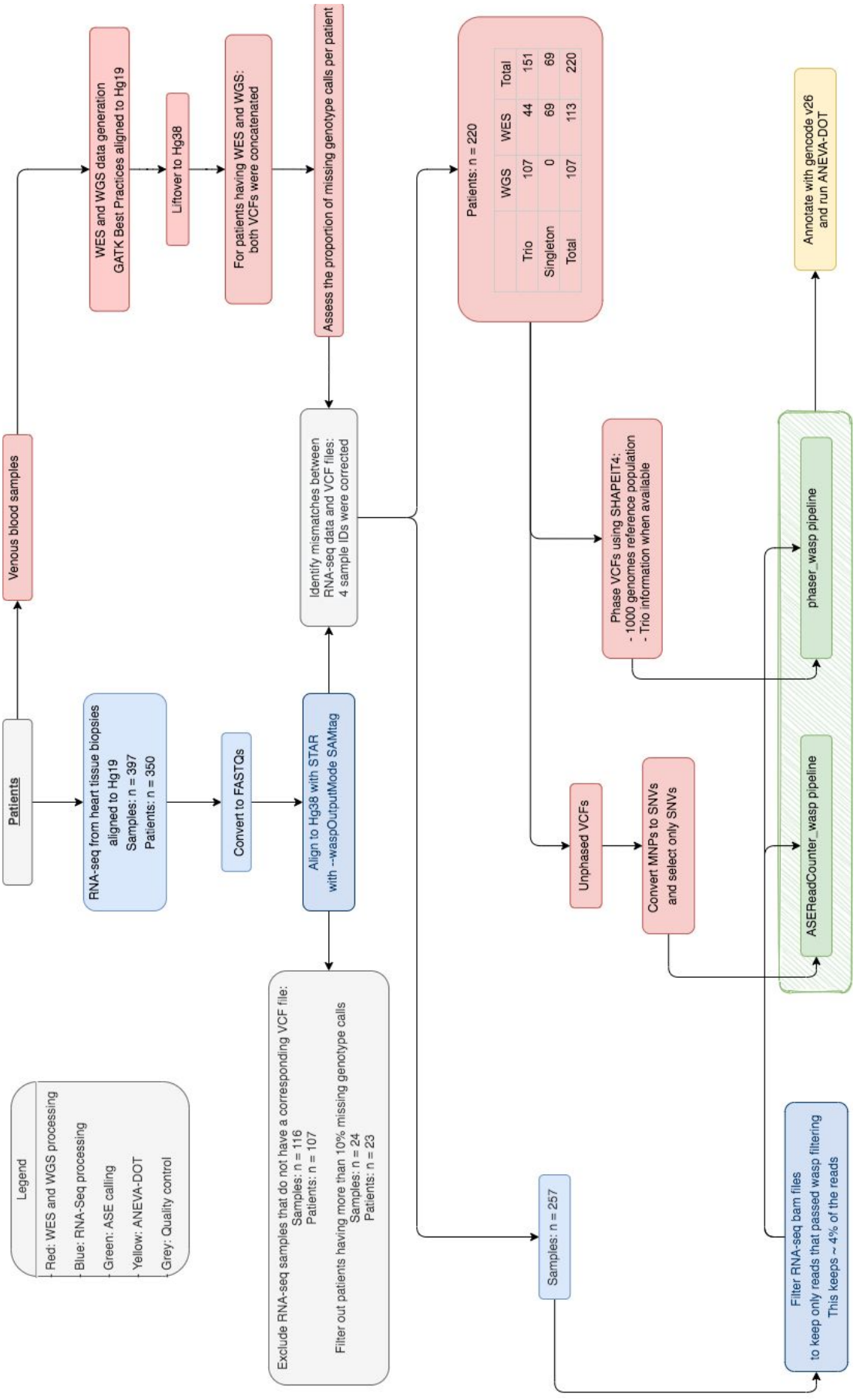
Even though our patients' VCF files had already been reviewed by a team of experts, annotation update and WGS / WES reanalysis has proven useful before (82–86). Each patient’s VCF file was annotated using Ensembl Variant Effect Predictor (VEP) v.99.2 (87) and dbNSFP4.1a (88,89).

We first assessed whether variants found within ANEVA-DOT’s outlier genes could be classified as pathogenic or likely pathogenic according to the ACMG-AMP standards and guidelines for the interpretation of sequence variants (47).

We then filtered variants to retain only variants found within genes previously implicated in Congenital Heart Diseases (hereafter: “panel genes”), that were rare (< 0.1% in gnomAD v3 (90)) and with genotype quality > 20. We assessed whether these variants could be classified as pathogenic or likely pathogenic according to the ACMG-AMP guidelines for each patient. Our list of panel genes was curated from various sources: ClinGen gene list of CHD genes updated by clingen experts working group (not yet released), heart disease

diagnostic panel from Nantes teaching hospital as well as panels from private genetic testing companies (Invitae “Congenital Heart Disease Panel”, Fulgent “Congenital Heart Defect NGS Panel”, Blueprints Genetics “Congenital Structural Heart Disease Panel”, CeGaT GmbH “Congenital Heart Defects Panel” and Fulgent “Comprehensive Cardiovascular NGS Panel”). In total, our list of panel genes included 396 unique genes (Supplementary List 1). For trios, we also assessed all de novo variants found within panel genes in accordance with the ACMG-AMP guidelines.

2.9 Flowchart overview



3. Results

3.1 Trisomy 21 patients show allelic imbalance on chromosome 21

15/220 (6.8%) patients had trisomy 21. We suspected that these patients would display strong allelic imbalance for chromosome 21 genes. Figure 2 shows that genes on chromosome 21 for these patients follow a ⅓ ⅔ expression pattern due to the additional chromosome 21 whereas the pattern displayed by genes on chromosomes other than chromosome 21 shows a balanced allelic expression (50/50 ratio).

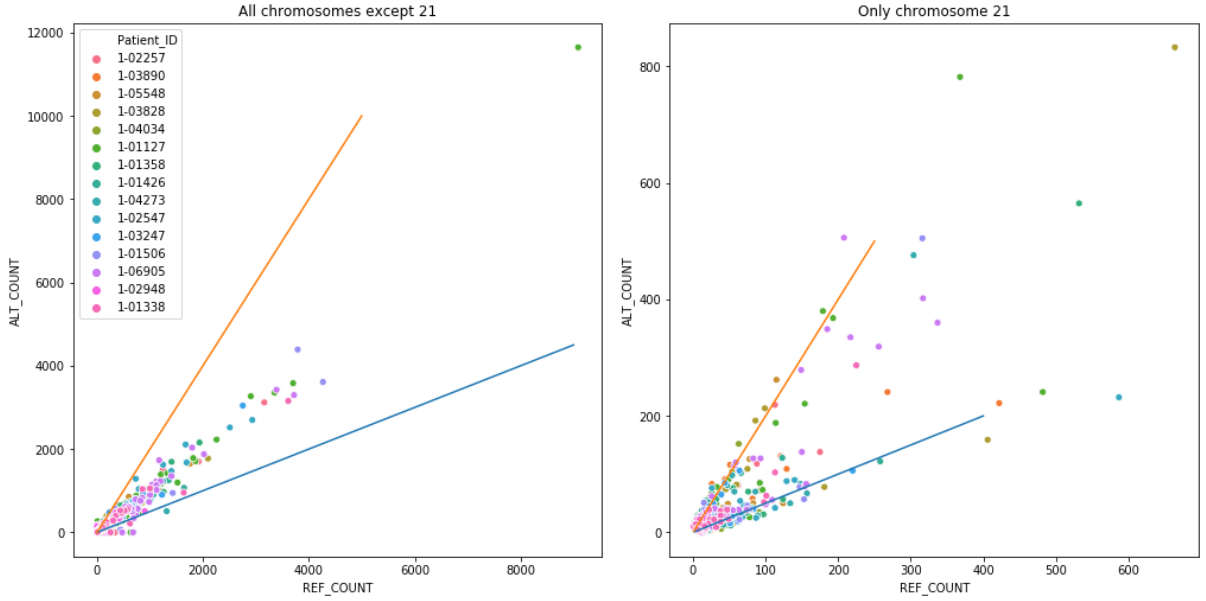


Figure 2: Comparison between reference and alternative counts for trisomy 21 patients.

For each gene that could be tested by ANEVA-DOT, comparison between reference and alternative counts for genes outside of chromosome 21 (left panel) and genes on chromosome 21 (right panel). Blue and orange lines represent the theoretical ⅓ and ⅔ distribution of allelic expression.

Figure 3 shows a strong enrichment in low p-values after ANEVA-DOT on chromosome 21 compared with other chromosomes of patients with trisomy 21. Overall, 9.2% of the genes on chromosome 21 were outliers as compared to a maximum of 1.4% for chromosomes other than chromosome 21. This difference was expected since ANEVA-DOT tests against

the null hypothesis that the observed allelic imbalance in an individual is consistent with dosage variation in the general population. However, ANEVA-DOT expects a 50/50 ratio as the null which is not a valid hypothesis for patients with trisomy 21 (⅓ ⅔ ratio of allelic expression).

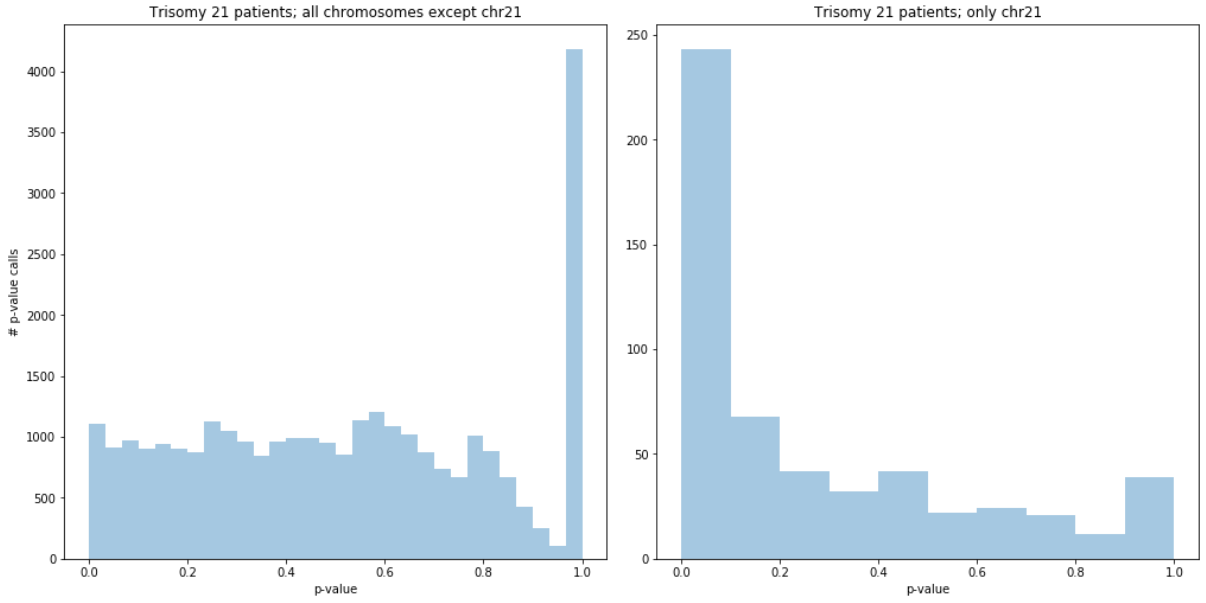


Figure 3: Distribution of ANEVA-DOT p-values for trisomy 21 patients.

Distribution of p-values for genes outside of chromosome 21 (left panel) compared to genes on chromosome 21 (right panel).

Because trisomy 21 patients show allelic imbalance on chromosome 21 that biases the ANEVA-DOT test, chromosome 21 data for patients with trisomy 21 was removed from all subsequent analysis.

3.2 phASER increases the number of genes tested while keeping outlier genes low

In order to assess each pipeline's ability to capture allelic imbalance signal in relevant genes, we computed the number of genes that could be tested by ANEVA-DOT, the number of outlier genes: genes that were significant at 5% false discovery rate (FDR) and the ratio of the number of [significant] / [tested] genes. Table 1 shows the number of genes tested and the number of outlier genes for the top 5 samples with the highest ratio [significant] / [tested].

One sample (1-05121_CHD-L233-1_DA) had a number of significant genes ~ 6 times higher than the second maximum number of outlier genes and was removed from all subsequent analysis.

Table 1: Number of genes tested, significant and ratio of the number of genes [significant] / [tested] using phASER pipeline for the top 5 samples with the highest ratio.

Sample_ID	Number of genes tested	Number of outlier genes	Ratio [significant] / [tested]
1-05121_CHD-L233-1_DA	2060	223	0.11
1-00842_CHD-C13-1_RA	1030	38	0.04
1-00842_CHD-C13-3_LA	1453	50	0.03
1-00842_CHD-C13-4_LV	1576	47	0.03
1-04693_CHD-L218-2_AS	3307	96	0.03

We then computed the number of unique genes that could be tested and the number of unique outlier genes across all samples for each tissue, and separated samples that had WES from samples that had WGS (Figure 4). On average, phASER was able to test 1.70 times more unique genes compared with ASEReadCounter while finding 0.69 times less unique outlier genes (Supplementary Table 1).

The ASEReadCounter pipeline was able to test 1.16 times more unique genes and find 1.84 times more unique outlier genes across samples if WGS was performed compared to WES on average. (Supplementary Table 2)

phASER pipeline was able to test 1.39 times more unique genes and find 1.32 times more unique outlier genes across samples if WGS was performed compared to WES on average. (Supplementary Table 2)

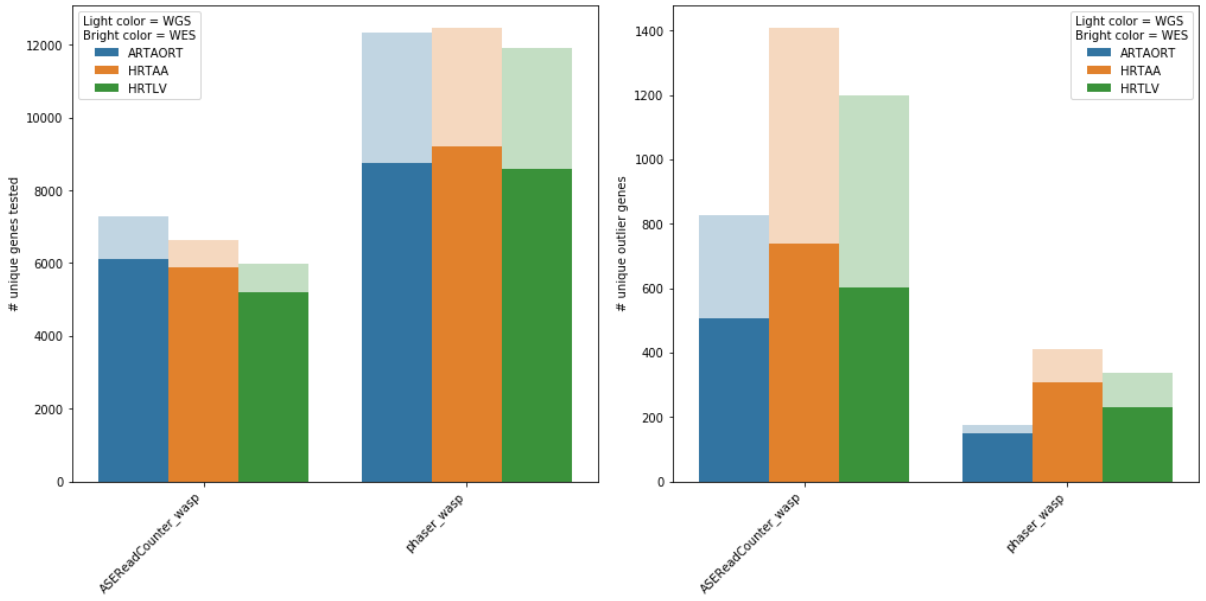


Figure 4: Number of unique genes tested and number of unique outlier genes across all samples.

Left panel: Number of unique genes that could be tested by ANEVA-DOT across all samples, according to the pipeline, the samples' tissue of origin and the sequencing method (bright colors correspond to WES and light colors correspond to WGS). Right panel: Number of outlier genes after ANEVA-DOT across all samples, according to the same categories as the left panel.

For patients who had **WGS** sequencing performed, the median number of genes **tested** per sample was 2835 and 5585 for ASEReadCounter_wasp and phaser_wasp pipelines, respectively (Figure 5).

For patients who had **WES** sequencing performed, the median number of genes **tested** per sample was 1467 and 1842 for ASEReadCounter_wasp and phaser_wasp pipelines, respectively (Figure 5).

For patients who had **WGS** sequencing performed, the median number of **outlier** genes per sample was 46 and 9 for ASEReadCounter_wasp and phaser_wasp pipelines, respectively (Figure 5).

For patients who had **WES** sequencing performed, the median number of **outlier** genes per sample was 28 and 7 for ASEReadCounter_wasp and phaser_wasp pipelines, respectively (Figure 5).

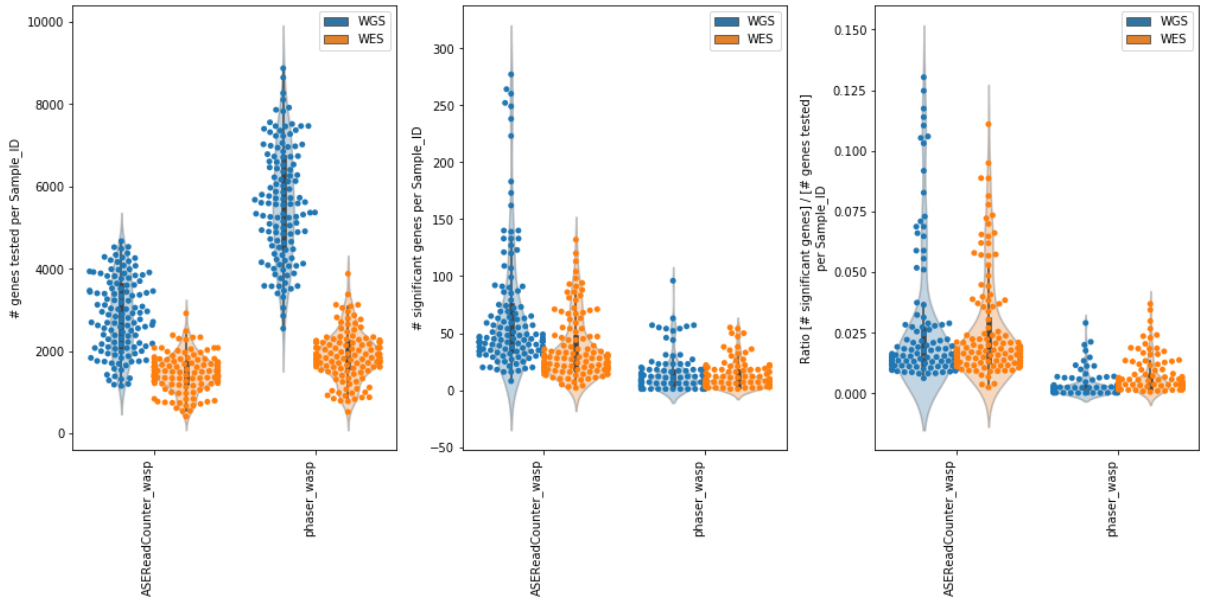


Figure 5: Number of genes tested, significant and ratio of [significant] / [tested] genes for each sample.

Each dot represents a sample for which we computed the number of genes tested (left panel), significant (middle panel) and the ratio of the two (right panel). Orange and blue colors represent samples for whom WES or WGS sequencing was performed, respectively.

For any given sample, ANEVA-DOT can only be calculated for the fraction of genes that have deep enough RNA-seq coverage and at least one heterozygous variant to distinguish between reference and alternate alleles during ASE data generation. Table 2 shows the median proportion of panel genes that could be tested by ANEVA-DOT per individual according to the pipeline used and the sequencing method. In the original ANEVA-DOT paper, Mohammadi *et al.* tested ANEVA-DOT in the general population of 466 skeletal muscle samples from GTEx. They describe a median of 3390 genes tested per individual and an average of 56% of the genes previously implicated in neuromuscular disorders to be testable per individual (58). This proportion is very similar to what we show for phASER

pipeline and WGS despite the fact that RNA-sequencing depth in the original muscle dataset was on average over 40% deeper than for our samples (91).

Table 2: Proportion of panel genes that could be tested by ANEVA-DOT according to the pipeline used and the sequencing method.

Pipeline	WES or WGS	Median proportion of panel genes tested
ASEReadCounter_wasp	WGS	35.6%
ASEReadCounter_wasp	WES	20.7%
phaser_wasp	WGS	56.3%
phaser_wasp	WES	23.8%

phASER allows for testing of a larger number of genes per sample and a larger proportion of panel genes compared with ASEReadCounter. Besides, WGS samples have about 3 times as many genes tested and twice as many panel genes tested compared with WES.

3.3 phASER increases coverage and lowers frequent outliers

For each gene tested across any sample, we compared the total coverage (REF count + ALT count) between ASEReadCounter and phASER pipelines for each tissue (Figure 6). Total coverage was highly positively correlated but phASER allows for better coverage with a median coverage 1.38 times higher if phASER is used.

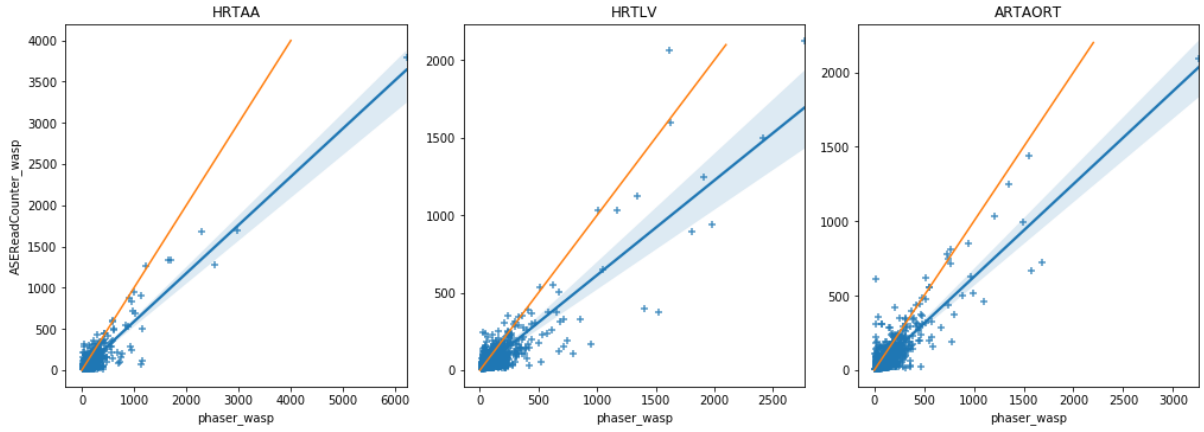


Figure 6: Gene coverage comparison between ASEReadCounter and phASER pipelines.

For each gene, comparison of its total coverage between phASER (x axis) and ASEReadCounter (y axis) for Atrial Appendage (left panel), Left Ventricle (middle panel), and Aorta (right panel) tissues. Orange lines represent the theoretical x = y distribution.

Some genes can appear as outlier genes in multiple samples. Frequent outlier genes are genes that appeared as outliers in > 1% of the samples. These genes likely represent rare situations where V^G calculated based on one allelic expression SNV (aeSNV) captures different regulatory variance patterns than other aeSNVs in the same gene that may be used for ANEVA-DOT analysis of a given individual (58). Some frequent outlier genes may also arise due to the fact that outlier status was associated with ancestry, sex, or sequencing platform that introduce a systematic bias in ANEVA-DOT calculation (58). However, since our cohort only includes patients presenting similar phenotypes, frequent outlier genes could also, in principle, include disease genes. For instance, *ELN*, a gene known to be implicated in supravalvular aortic stenosis (92,93) and *SMAD6*, known to be implicated in aortic isthmus stenosis and bicuspid aortic valve (94–96) are genes that appear as frequent outliers in our cohort (5.5%, 95% confidence interval [CI]: [1.1%-15%] and 8.3%, 95% CI: [1.0%-27%], respectively).

Mohammadi *et al.* showed that in the general population of 466 skeletal muscle samples from GTEx (58) 1.93% of the genes tested were frequent outliers (113 out of 5848 genes

tested) which is less than our ASEReadCounter pipeline (Table 3). However, ANEVA-DOT was developed using the GTEx dataset which may reduce the noise when testing the method on this same dataset. Besides, V^G scores originate from the GTEx samples which were sequenced at a greater depth and include healthy adults whereas our population is composed of newborn patients for the main part which may not be fully comparable (97,98). Our phASER pipeline shows less frequent outliers than ASEReadCounter and less frequent outliers than the original method.

Table 3: Absolute counts and proportion of frequent outlier genes.

Pipeline	Tissue	Number of genes not frequent outlier	Number of frequent outliers genes	Proportion of frequent outliers
ASEReadCounter_wasp	ARTAORT	7025	287	3.93%
ASEReadCounter_wasp	HRTAA	6261	386	5.81%
ASEReadCounter_wasp	HRTLV	5503	468	7.84%
phaser_wasp	ARTAORT	12400	35	0.28%
phaser_wasp	HRTAA	12483	53	0.42%
phaser_wasp	HRTLV	11903	85	0.71%

Since phaser_wasp pipeline allows for better coverage, a higher number of genes tested, a lower number of outlier genes per sample, a lower proportion of frequent outlier genes and a higher proportion of panel genes tested, all subsequent analyses were made based on phaser_wasp pipeline data.

3.4 Outlier genes are enriched for rare and deleterious variants

We computed the relative risk ratio for an outlier gene to carry a rare heterozygous variant (minor allele frequency < 1%) in a 10-kb window upstream of the transcriptional start site and in the gene body compared with non-outlier genes. We show that the lower the p-value is at 5% FDR, the more enriched for rare variants an outlier gene is, compared to non-outlier genes (Figure 7). The weaker signal for WGS data may be due to the fact that introns are captured better in WGS data and contain many rare variants but only a few are functional (Supplementary Table 3).

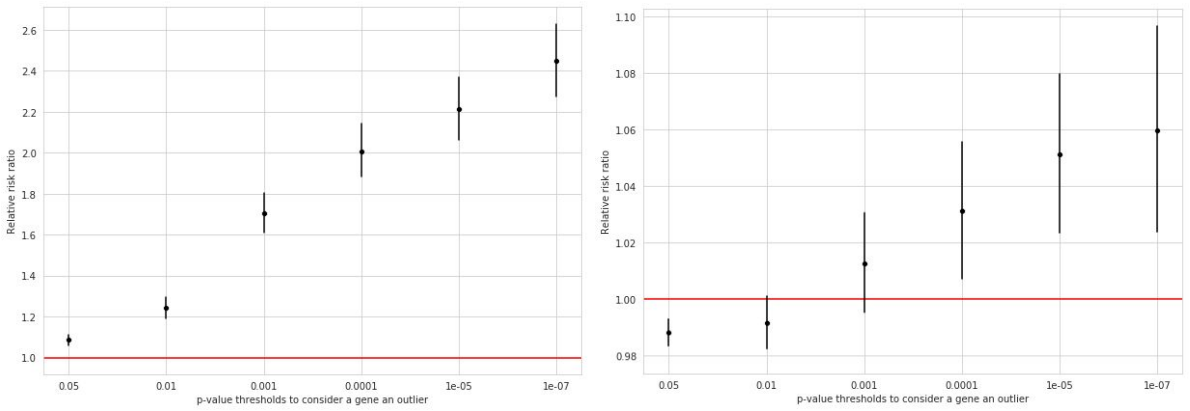


Figure 7: Enrichment for rare variants (minor allele frequency < 1%) in outlier genes as a function of outlier p-value.

Left panel: Relative risk ratio for an outlier gene to carry a rare variant when considering patients for whom WES was performed. Right panel: Relative risk ratio for an outlier gene to carry a rare variant when considering patients for whom WGS was performed.

This enrichment was particularly pronounced for putative rare (minor allele frequency < 1%) gene-disrupting variants that are expected to have a strong effect on gene expression levels by nonsense-mediated decay (Figure 8). The top 3 most enriched consequence categories are splice donor variants, splice acceptor variants and stop gained variants for WGS patients with a relative risk ratio of 3.9, 3.6 and 3.1, respectively and protein altering variants, stop gained variants and splice donor variants for WES with a relative risk ratio of 9.4, 7.7 and 4.0, respectively.

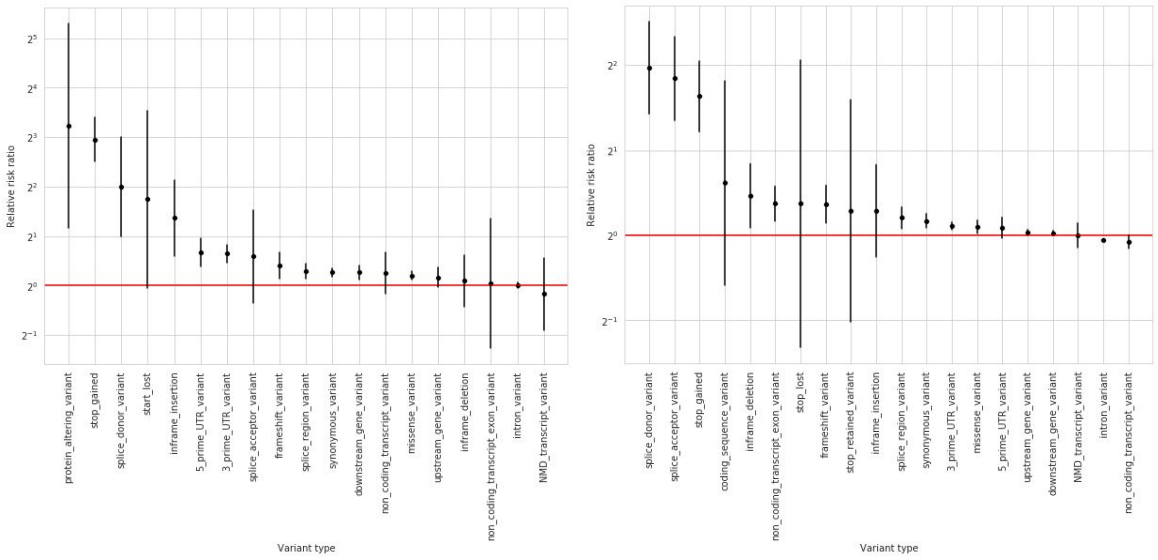


Figure 8: Enrichment for rare (minor allele frequency < 1%) gene-disrupting variants in outlier genes according to their predicted consequence.

Left panel: Relative risk ratio for an outlier gene to carry a rare gene-disrupting variant when considering patients for whom WES was performed. Right panel: Relative risk ratio for an outlier gene to carry a rare gene-disrupting variant when considering patients for whom WGS was performed.

3.5 Previously resolved cases

After careful evaluation of patients variants following the ACMG-AMP guidelines by a team of experts from New York-Presbyterian Hospital/Columbia University Medical Center Clinical Genetics and Genomics department, diagnosis were made for 7/220 patients. Two patients carried microdeletion / microduplication which can not be spotted using ANEVA-DOT since outlier genes should carry variants that lead to allelic imbalance to be detected: typically loss-of-function or essential splice site variants (52). Patients 1-01026, 1-02038 and 1-02838 incriminated variants were: *NOTCH1*(NP_060087.3:p.Asn510LysfsTer2), *CHD7* (NP_060250.2:p.Tyr2601Ter) and *KMT2D* (NP_003473.3:p.Gln1893Ter), respectively. These genes could not be tested by ANEVA-DOT for these patients due to low coverage. Patients 1-00541 and 1-00596 incriminated variants were *RAI1*(ENSP00000463984.1:p.Ala84fs, non-canonical transcript) and *KMT2D* (NP_003473.3:p.Ser1722ArgfsTer9), respectively. These genes were tested by

ANEVA-DOT but no allelic imbalance could be detected. No other strong candidate was found for these patients.

3.6 Heart relevant pathogenic variants

Patient 1-05368 is a newborn patient presenting with coarctation of the aorta, atrial septal defect, double outlet right ventricle and hypoplastic left ventricle. ANEVA-DOT identified RNA Binding Fox-1 Homolog 2 (*RBFOX2*) as the third most significant outlier gene for this patient, displaying complete allelic imbalance. The patient carries a de novo, rare (Allelic Count (AC) = 1 in gnomAD v3), heterozygous, stop gain variant (NM_001082579.2:c.1069C>T, NP_001076048.1: p.Arg357Ter) in *RBFOX2.* This gene is highly intolerant to loss of function variants with an observed / expected (o/e) number of loss-of-function variants ratio of 0.04 (90% CI: 0.01 - 0.19). This variant is pathogenic according to the ACMG-AMP guidelines.

RBFOX2 encodes an RNA binding protein that is thought to be a key regulator of alternative exon splicing in the nervous system and other cell types. The protein binds to a conserved UGCAUG element found downstream of many alternatively spliced exons and promotes inclusion of the alternative exon in mature transcripts (99). Rbfox2 is critical for zebrafish heart development (100) and regulates epithelial-mesenchymal transitions (101) which is thought to underlie hypoplastic left heart syndrome pathogenesis (102).

Even though *RBFOX2* is not in our curated list of panel genes, deleterious Single Nucleotide Variants (SNVs) in this gene have been implicated in CHD and more specifically hypoplastic left heart syndrome (103,104). In 2014, Glessner *et al.* also reported a de novo copy number loss that encompasses *RBFOX2* in a CHD proband with hypoplastic left heart syndrome (105).

Given the segregation evidence, the phenotype match and the allelic imbalance, it is reasonable to think that this gene's disruption is responsible for the patient's disease.

Patient 1-02627 is an 8 years old patient presenting with tetralogy of Fallot, strabismus, hearing loss and cleft lip and palate. For this patient, ANEVA-DOT prioritized the endoglin

(*ENG)* gene (p-value of 2.8e-10 at 5% FDR). *ENG* is highly intolerant to loss-of-function variants with an o/e ratio of 0.09 (90% CI: 0.04 - 0.24) and found within Blueprints Genetics "Congenital Structural Heart Disease Panel". The patient carries a de novo, heterozygous, frameshift variant (NM_001114753.2: c.742_743insT, NP_001108225.1:p.Asp248ValfsTer86) unknown from gnomAD v3 and displays strong allelic imbalance based on RNA-seq data. This variant is pathogenic according to the ACMG-AMP guidelines.

ENG encodes a homodimeric transmembrane protein which is a major glycoprotein of the vascular endothelium as part of the transforming growth factor beta (TGF-β) receptor complex. Endoglin loss of function endothelial cells were shown to acquire venous characteristics and displayed secondary endoglin-independent proliferation resulting in arteriovenous malformation (AVM) (106). Besides, endoglin mutant blood vessels were shown to continue to enlarge in response to flow increases, exacerbating pre-existing embryonic arterial-venous shunts (107). This gene is also involved in heart development (108). Deleterious variants in this gene cause hereditary hemorrhagic telangiectasia, an autosomal dominant multisystemic vascular dysplasia. Further variant interpretation is required to assess the causal relationship between the frameshift variant and the observed cardiac phenotype. This variant may also be considered an incidental finding, however *ENG* is not in the ACMG list of genes recommended for reporting incidental findings (109,110).

For patient 1-06936, a 3 months old patient presenting with ventricular septal defect, ANEVA-DOT prioritized Mitogen-Activated Protein Kinase Kinase 1 (*MAP2K1*) (p-value 2.5e-05 at 5% FDR) which displayed complete allelic imbalance. The patient carries a heterozygous, stop gain variant (NM_002755.3:c.741G>A, NP_002746.1:p.Trp247Ter), unknown from gnomAD v3, inherited from his mother. This variant is pathogenic according to the ACMG-AMP guidelines. MAP2K1 is intolerant to loss of function variants with an o/e ratio of 0.15 (90% CI: 0.07 - 0.38).

MAP2K1 encodes a protein kinase that lies upstream of MAP kinases and stimulates the enzymatic activity of MAP kinases upon a wide variety of extra- and intracellular signals. As

an essential component of MAP kinase signal transduction pathway, this kinase is involved in many cellular processes. Diseases associated with MAP2K1 include cardiofaciocutaneous syndrome 3, an autosomal dominant disorder involving characteristic craniofacial features, hair and skin abnormalities, postnatal growth deficiency, hypotonia, developmental delay and cardiovascular malformations including ventricular septal defects (111). Distinctive features of *MAP2K1* variant-positive cases include macrostomia and horizontal shape of palpebral fissures (112). However, these lesions typically develop over time and thus may not be helpful given our patient's age. Further analyses are needed to assess a potential incomplete penetrance mechanism for this variant.

3.7 Deleterious variants

Patient 1-01021 was 10 months old at the time of enrollment. Even though clinical evaluation of extra-cardiac manifestation can be difficult given the patient's age, the only reported phenotype for this patient was D-loop transposition of the great arteries with ventricular septal defect and left ventricular outflow tract obstruction as well as laryngomalacia. The Phosphoglucomutase 1 (*PGM1*) gene was identified by ANEVA-DOT as an outlier gene (p-value 1.5e-5 at 5% FDR) and displayed complete allelic imbalance. The patient carries a stop gain, heterozygous variant (NM_002633.3:c.763G>T, NP_002624.2:p.Glu255Ter), unknown from gnomAD v3, in *PGM1*. Parents' genotypes were not available.

There are several PGM isozymes, which are encoded by different genes and catalyze the transfer of phosphate between the 1 and 6 positions of glucose. In most cell types, PGM1 is predominant, representing about 90% of total PGM activity. Decreased PGM1 activity is responsible for congenital disorder of glycosylation type It and glycogen storage disease inherited on an autosomal recessive mode. This disorder encompasses a wide range of clinical manifestations, most commonly presenting with Pierre Robin sequence, bifid uvula with or without cleft palate at birth, growth delay, chronic hepatitis, exercise-related fatigue, muscle weakness, intermittent hypoglycemia, and dilated cardiomyopathy and/or cardiac arrest and has been implicated in severe familial cardiomyopathy (113–115). Further

investigations are needed to clarify this variant's inheritance, find a potential second hit and assess overall involvement in the patient's phenotype

Patient 1-00384 is a 17 years old female of african ancestry presenting with hypoplastic left heart syndrome and mitral atresia. ANEVA-DOT identified Alanyl-TRNA Synthetase Domain Containing 1 (*AARSD1*) as an outlier gene in the 3 tissues available for this patient. A rare (AC = 38 and Allele Frequency (AF) = 0.0009 in gnomAD v3 for the African population) loss of function, heterozygous splice site variant (NM_001261434.2: c.1104-1G>T) inherited from her mother was identified in *AARSD1*. This variant is of uncertain significance according to the ACMG-AMP guidelines. *AARSD1* is expected to be tolerant to loss of function variants with an o/e ratio of 0.73 (90% CI: 0.49 - 1.11).

AARSD1 encodes a muscle-specific Hsp90 cochaperone whose knock down was shown to interfere with the differentiation of myoblasts into myotubes (116). More specifically, *AARSD1* is part of a complex that promotes heme insertion into immature apo-Myoglobin, and thus generates functional myoglobin during muscle myotube formation (117).

Readthrough transcription between the neighboring *PTGES3L* (prostaglandin E synthase 3 (cytosolic)-like) and *AARSD1* genes also occurs at this locus. The readthrough transcript encodes a fusion protein that is also affected by this variant (NM_001136042.2: c.1626-1G>T). The function of the fusion protein remains unclear.

This variant's impact on splicing remains to be clarified (Figure 9). Further analyses are needed to assess the relevance of this finding and a potential incomplete penetrance mechanism for the mother.

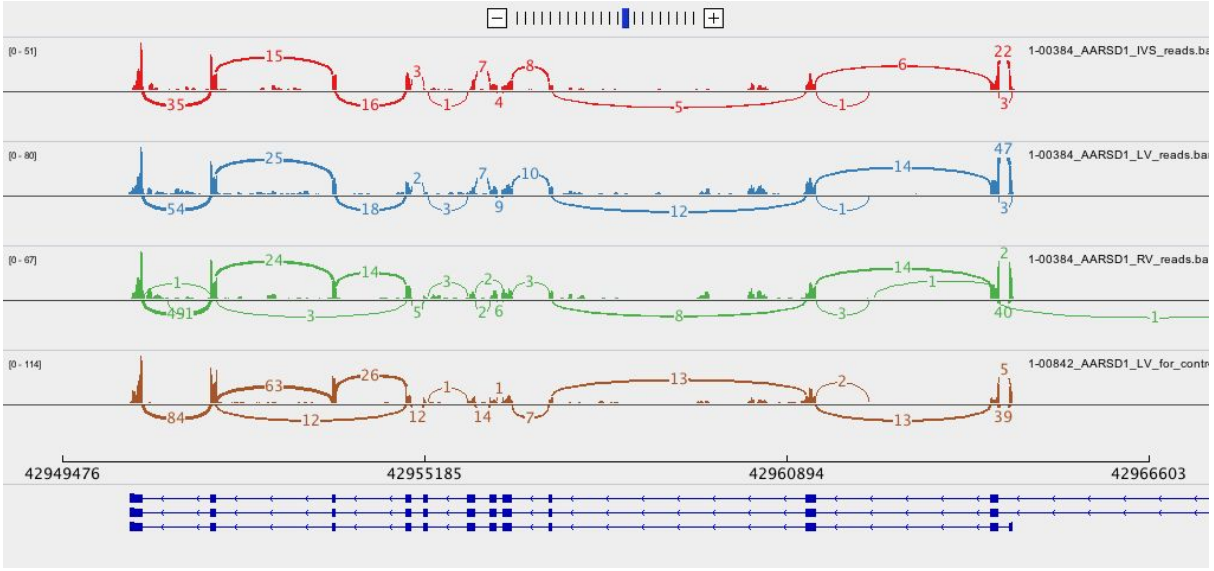


Figure 9: Sashimi plot for *AARSD1* for all available samples from patient 1-00384.

Patient 1-00842, also of african ancestry, does not carry any suspicious variant in this gene and was added for comparison purposes.

3.8 Are frequent outlier genes relevant?

Two genes, elastin (*ELN*) and SMAD Family Member 6 (*SMAD6*), appeared as frequent outliers in our cohort (5.5%, 95% confidence interval [CI]: [1.1%-15%] and 8.3%, 95% CI: [1.0%-27%], respectively) but also belonged to gene panels. We assess below the relevance of the variants found within these genes.

ELN encodes a protein that is one of the two components of elastic fibers. Degradation products of the encoded protein, known as elastokines, bind the elastin receptor complex and other receptors and stimulate migration and proliferation of monocytes and skin fibroblasts. Elastokines can also contribute to cancer progression. Deletions and SNVs in this gene are associated with supravalvular aortic stenosis and autosomal dominant cutis laxa (92,93). This gene was an ANEVA-DOT outlier for the aorta tissue type in 3 patients out of 55 tested. Patient 1-02969 presented with hypoplastic left heart syndrome and only carried 2 intronic variants in this gene, both inherited from his father. Patient 1-05476 presented with tetralogy of Fallot and only carried intronic variants in this gene. One homozygous for the alternate allele inherited from his heterozygous mother and

heterozygous father and another intronic de novo variant. No splicing alteration could be seen in the *ELN* gene for this patient. Patient 1-05672 presented with coarctation of the aorta and had an intronic variant in *ELN*, inherited from his mother with AF > 0.1% in gnomAD v3 for europeans.

SMAD6 encodes a protein that functions in the negative regulation of bone morphogenetic proteins (BMP) and TGF-β/activin-signalling and is known to be implicated in aortic isthmus stenosis and bicuspid aortic valve (94–96). This gene was an ANEVA-DOT outlier for the Aorta tissue type in 2 patients out of 24 tested. Patient 1-06516 presented with tetralogy of Fallot and carried 1 rare intronic variant inherited from the patient's father. No splicing alteration could be seen in *SMAD6* for this patient. Patient 1-07300 presented with VACTERLS syndrome and carried one intronic rare variant and one intronic variant with AF > 0.1% in gnomAD v3, inherited from each parent as well as a rare downstream gene variant inherited from his father but no impact on *SMAD6* expression was found (Table 4).

Table 4: Frequent outlier genes variants

Patient	Gene	Phenotype	Variant	Genotypes		
				Proband	Mother	Father
1-02969	*ELN*	Hypoplastic left heart syndrome	ENST00000358929.8:c.1097-333G>A	0/1	0/0	0/1
			ENST00000358929.8:c.2044+32G>A	0/1	0/0	0/1
1-05476	*ELN*	Tetralogy of Fallot	ENST00000358929.8:c.572-204A>C	1/1	0/1	0/1
			ENST00000358929.8:c.800-1420T>C	0/1	0/0	0/0
1-05672	*ELN*	Coarctation of the aorta	ENST00000358929.8:c.2179+736T>A	0/1	0/1	0/0
1-06516	*SMAD6*	Tetralogy of Fallot	ENST00000288840.10:c.952+8582G>A	0/1	0/0	0/1
1-07300	*SMAD6*	VACTERLS syndrome	ENST00000288840.10:c.952+7252A>G	0/1	0/1	0/0
			ENST00000288840.10:c.953-22801G>A	0/1	0/0	0/1
			downstream_gene_variant	0/1	0/0	0/1

Even though *SMAD6* and *ELN* are genes previously known to be implicated in congenital heart diseases, phenotypic findings in patients from our cohort carrying variants in these

genes do not seem to match previous descriptions and variants found in these genes do not seem relevant to our patients' conditions. Their outlier status seems artifactual.

4. Discussion

phASER - ANEVA-DOT is more accurate than previous methods. In this study, we introduce the phASER - ANEVA-DOT pipeline to compare individual transcriptome to previously generated reference data at haplotype level. We show that using phASER for ASE data generation allows for testing more genes than using ASEReadCounter while finding less outlier genes making this method a fast and powerful approach for finding genes carrying variants with likely deleterious effects. WGS largely increases the number of genes tested compared to WES and using phASER keeps the number of outlier genes low so that even at this scale, the small numbers of outliers make further manual curation feasible in a clinical setting without compromising sensitivity. Besides, in 2016, McKean *et al.* (72) employed RNA-seq in a cohort of CHD subjects partially overlapping with our cohort. They specifically assessed ASE genes with loss of function variants in the silenced allele for 7 patients also included in our cohort. They argue that only one of these genes, *RBFOX2*, should be considered a CHD gene candidate while the others are probably artifactual. 4/7 genes from McKean *et al.* study were ANEVA-DOT outliers including *RBFOX2*, *AARSD1* and *PGM1* for patients 1-05368, 1-00384, 1-01021, respectively, while the other 3 genes were not. This finding highlights that ANEVA-DOT may be more accurate for gene prioritization than previous methods.

As suggested before (58), we confirm that frequent outlier genes in the context of large cohorts of rare heterogeneous diseases are probably irrelevant.

This method is effective for gene-prioritization. ANEVA-DOT captures transcriptome outcomes of genetic effects without having to identify rare regulatory variants themselves (58). This enables the identification of genes where a patient carries a heterozygous variant with an unusually strong effect on gene expression. We show that this approach is effective in our CHD cohort leading to one new probable diagnosis in this cohort of mostly unresolved cases so far. We identify *RBFOX2* as a candidate CHD gene and argue that this gene should be included in CHD routine gene panels.

Despite these advantages, our method has several limitations. The main caveats are that ANEVA-DOT relies on V^G estimates, requires an aeSNV to generate ASE data from and high enough RNA-seq coverage to be able to perform the Dosage Outlier Test. V^G may be lacking or noisy for genes with few coding variants owing to small size, high coding constraint, or low expression levels (58). High coverage RNA-seq data coming from disease-relevant tissues is necessary but technical issues or pathogenic biallelic loss of expression in genes can make ANEVA-DOT inapplicable. These limitations make ANEVA-DOT only applicable to about half of expressed genes per sample.

On the patients' side, identifying the specific variants underlying outliers is still challenging, especially for noncoding regions. Congenital heart diseases are complex and involve many genes and mechanisms. Also, tissue samples are usually obtained after birth but diseases may result from in utero gene disruption. Last, ASE *per se* may not be enough to cause a disease phenotype, particularly if dosage compensation restores this gene's function.

5. Conclusion

Like other genetic diagnosis tools, phASER - ANEVA-DOT should be used in conjunction with other methods to capture different types of rare variants underlying diseases. With the increasing availability of transcriptome and genomic data in a clinical setting phASER - ANEVA-DOT proves fast and accurate to become a powerful additional layer to integrate these data and interpret the genome and its disease-contributing variants.

6. Bibliography

1. Schwarze K, Buchanan J, Fermont JM, Dreau H, Tilley MW, Taylor JM, et al. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. Genet Med. 2020;22(1):85‑94. DOI: 10.1038/s41436-019-0618-7
2. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies: Overview of Next-Generation Sequencing. Curr Protoc Mol Biol. 2018;122(1):e59. DOI: 10.1002/cpmb.59
3. Pereira MA, Malta FSV, Freire MCM, Couto PGP. Application of Next-Generation Sequencing in the Era of Precision Medicine. Dans: Application of Next-Generation Sequencing in the Era of Precision Medicine. INTECH; 2017.
4. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418‑26. DOI: 10.1016/j.tig.2014.07.001
5. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. Nature Publishing Group; 2020;577(7789):179‑89. DOI: 10.1038/s41586-019-1879-7
6. Xue Y, Ankala A, Wilcox WR, Hegde MR. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. Genet Med. 2015;17(6):444‑51. DOI: 10.1038/gim.2014.122
7. LaDuca H, Stuenkel A, Dolinsky JS, Keiles S, Tandy S, Pesaran T, et al. Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients. Genet Med. 2014;16(11):8. DOI: 10.1038/gim.2014.40
8. Beltran H, Eng K, Mosquera JM, Sigaras A, Romanel A, Rennert H, et al. Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. JAMA Oncol. 2015;1(4):466‑74. DOI: 10.1001/jamaoncol.2015.1313
9. Rusch M, Nakitandwe J, Shurtleff S, Newman S, Zhang Z, Edmonson MN, et al. Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. Nat Commun. 2018;9(1):3962. DOI: 10.1038/s41467-018-06485-7
10. Leggett RM, Ramirez-Gonzalez RH, Clavijo BJ, Waite D, Davey RP. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Front Genet. 2013;4(288):5. DOI: 0.3389/fgene.2013.00288
11. Ekblom R, Wolf JBW. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014;7(9):1026‑42. DOI: 10.1111/eva.12178
12. Chrystoja CC, Diamandis EP. Whole Genome Sequencing as a Diagnostic Test: Challenges and Opportunities. Clin Chem. 2014;60(5):10. DOI: 10.1373/clinchem.2013.209213
13. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57‑63. DOI: 10.1038/nrg2484
14. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17:15. DOI: 10.1038/nrg.2016.10
15. Koch L. Exploring human genomic diversity with gnomAD. Nat Rev Genet. 2020;21(8):448‑448. DOI: 10.1038/s41576-020-0255-7
16. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2004;33:4. DOI: 10.1093/nar/gki033
17. Rath A, Olry A, Dhombres F, Maja MB, Urbero B, Ayme S. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. Hum Mutat. 2012;33(5):7. DOI: 10.1002/humu.22078
18. Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, et al. Whole-Genome Sequencing for Optimized Patient Management. Sci Transl Med. 2011;3(87):87re3-87re3. DOI: 10.1126/scitranslmed.3002243
19. Saunders CJ, Miller NA, Soden SE, Dinwiddie L, Noll A, Alnadi NA, et al. Rapid

Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. Sci Transl Med. 2015;3(4):27. DOI: 10.1126/scitranslmed.3004041
20. Liew WKM, Ben-Omran T, Darras BT, Prabhu SP, Yang Y, Eng CM, et al. Clinical Application of Whole-Exome Sequencing. JAMA Neurol. 2013;70(6):4. DOI: 10.1001/jamaneurol.2013.247
21. Biesecker LG, Green RC. Diagnostic Clinical Genome and Exome Sequencing. N Engl J Med. 2014;370(25):2418‐25. DOI: 10.1056/NEJMra1312543
22. Jamuar SS, Tan E-C. Clinical application of next-generation sequencing for Mendelian diseases. Hum Genomics. 2015;9(10):6. DOI: 10.1186/s40246-015-0031-5
23. Massalska D, Zimowski JG, Bijok J, Pawelec M, Czubak-Barlik M, Jakiel G, et al. First trimester pregnancy loss: Clinical implications of genetic testing: Pregnancy loss and genetic testing. J Obstet Gynaecol Res. 2017;43(1):23‐9. DOI: 10.1111/jog.13179
24. Mary E. N, Marc J. Clinical Management Guidelines For Obstetrician-Gynecologists. Number 43, May 2003: Management of Preterm Labor. Obstet Gynecol. 2003;101(5, Part 1):1039‐47. DOI: 10.1097/00006250-200305000-00052
25. Hanninen M, Klein GJ, Laksman Z, Conacher SS, Skanes AC, Yee R, et al. Reduced Uptake of Family Screening in Genotype-Negative Versus Genotype-Positive Long QT Syndrome. J Genet Couns. 2015;24(4):558‐64. DOI: 10.1007/s10897-014-9776-6
26. Theilade J, Kanters J, Henriksen FL, Gilså-Hansen M, Svendsen JH, Eschen O, et al. Cascade Screening in Families with Inherited Cardiac Diseases Driven by Cardiologists: Feasibility and Nationwide Outcome in Long QT Syndrome. Cardiology. 2013;126(2):131‐7. DOI: 10.1159/000350825
27. Hammill AM, Wentzel M, Gupta A, Nelson S, Lucky A, Elluru R, et al. Sirolimus for the treatment of complicated vascular anomalies in children. Pediatr Blood Cancer. 2011;57(6):1018‐24. DOI: 10.1002/pbc.23124
28. Iacobas I, Burrows PE, Adams DM, Sutton VR, Hollier LH, Chintagumpala MM. Oral rapamycin in the treatment of patients with hamartoma syndromes and PTEN mutation: Rapamycin for PTEN Mutation Hamartomas. Pediatr Blood Cancer. 2011;57(2):321‐3. DOI: 10.1002/pbc.23098
29. Williams K, Carson J, Lo C. Genetics of Congenital Heart Disease. Biomolecules. 2019;9(12). DOI: 10.3390/biom9120879
30. Lytzen R, Vejlstrup N, Bjerre J, Petersen OB, Leenskjold S, Dodd JK, et al. Live-Born Major Congenital Heart Disease in Denmark: Incidence, Detection Rate, and Termination of Pregnancy Rate From 1996 to 2013. JAMA Cardiol. 2018;3(9):829. DOI: 10.1001/jamacardio.2018.2009
31. Jenkins KJ, Correa A, Feinstein JA, Botto L, Britt AE, Daniels SR, et al. Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. Circulation. 2007;115(23):2995‐3014. DOI: 10.1161/CIRCULATIONAHA.106.183216
32. Øyen N, Poulsen G, Boyd HA, Wohlfahrt J, Jensen PKA, Melbye M. Recurrence of congenital heart defects in families. Circulation. 2009;120(4):295‐301. DOI: 10.1161/CIRCULATIONAHA.109.857987
33. Pediatric Cardiac Genomics Consortium. The Congenital Heart Disease Genetic Network Study. Circ Res. 2013;127:9. DOI: 10.1161/CIRCRESAHA.111.300297
34. Garg V, Kathiriya IS, Barnes R, Schluterman MK, King IN, Butler CA, et al. GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. Nature. 2003;424(6947):443‐7. DOI: 10.1038/nature01827
35. McDermott DA, Bressan MC, He J, Lee JS, Aftimos S, Brueckner M, et al. TBX5 Genetic Testing Validates Strict Clinical Criteria for Holt-Oram Syndrome. Pediatr Res. Nature Publishing Group; 2005;58(5):981‐6. DOI: 10.1203/01.PDR.0000182593.95441.64
36. Bedard JEJ, Haaning AM, Ware SM. Identification of a Novel ZIC3 Isoform and Mutation Screening in Patients with Heterotaxy and Congenital Heart Disease. PLoS ONE. 2011;6(8). DOI: 10.1371/journal.pone.0023755
37. Blue GM, Kirk EP, Giannoulatou E, Dunwoodie SL, Ho JWK, Hilton DCK, et al. Targeted Next-Generation Sequencing Identifies Pathogenic Variants in Familial

Congenital Heart Disease. J Am Coll Cardiol. 2014;64(23):2498-506. DOI: 10.1016/j.jacc.2014.09.048
38. Bonachea EM, Zender G, White P, Corsmeier D, Newsom D, Fitzgerald-Butt S, et al. Use of a targeted, combinatorial next-generation sequencing approach for the study of bicuspid aortic valve. BMC Med Genomics. 2014;7:56. DOI: 10.1186/1755-8794-7-56
39. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature. 2013;498(7453):220-3. DOI: 10.1038/nature12141
40. Vissers LELM, van Ravenswaaij CMA, Admiraal R, Hurst JA, de Vries BBA, Janssen IM, et al. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. Nat Genet. Nature Publishing Group; 2004;36(9):955-7. DOI: 10.1038/ng1407
41. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. Nature Publishing Group; 2010;42(9):790-3. DOI: 10.1038/ng.646
42. Kulkarni P, Frommolt P. Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. Comput Struct Biotechnol J. 2017;15:7. DOI: 10.1016/j.csbj.2017.10.001
43. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci. 2015;112(17):5473-8. DOI: 10.1073/pnas.1418631112
44. Hwang K-B, Lee I-H, Li H, Won D-G, Hernandez-Ferrer C, Negron JA, et al. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. Sci Rep. 2019;9(1):3219. DOI: 10.1038/s41598-019-39108-2
45. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169(7):1177-86. DOI: 10.1016/j.cell.2017.05.038
46. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102-10. DOI: 10.1093/hmg/ddv259
47. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-23. DOI: 10.1038/gim.2015.30
48. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017;18(10):599-612. DOI: 10.1038/nrg.2017.52
49. Magri F, Del Bo R, D'Angelo MG, Govoni A, Ghezzi S, Gandossini S, et al. Clinical and molecular characterization of a cohort of patients with novel nucleotide alterations of the Dystrophin gene detected by direct sequencing. BMC Med Genet. 2011;12:37. DOI: 10.1186/1471-2350-12-37
50. La Cognata V, D'Agata V, Cavalcanti F, Cavallaro S. Splicing: is there an alternative contribution to Parkinson's disease? neurogenetics. 2015;16(4):245-63. DOI: 10.1007/s10048-015-0449-x
51. Liu F, Gong C-X. Tau exon 10 alternative splicing and tauopathies. Mol Neurodegener. 2008;3:8. DOI: 10.1186/1750-1326-3-8
52. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9(386):eaal5209. DOI: 10.1126/scitranslmed.aal5209
53. Chen J, Weiss WA. Alternative splicing in cancer: implications for biology and therapy. Oncogene. Nature Publishing Group; 2015;34(1):1-14. DOI: 10.1038/onc.2013.570
54. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biol. 2015;16(1):195. DOI: 10.1186/s13059-015-0762-6
55. Zeng H, Edwards MD, Guo Y, Gifford DK. Accurate eQTL prioritization with an ensemble-based framework. Hum Mutat. 2017;38(9):1259-65. DOI: 10.1002/humu.23198
56. Brandt M, Gokden A, Ziosi M, Lappalainen T. A polyclonal allelic expression assay for detecting regulatory effects of transcript variants. Genome Med. 2020;12(1):79. DOI:

10.1186/s13073-020-00777-8
57. Richards J, Rivadeneira F, Inouye M, Pastinen T, Soranzo N, Wilson S, et al. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. Lancet. 2008;371(9623):1505‑12. DOI: 10.1016/S0140-6736(08)60599-1
58. Mohammadi P, Castel SE, Cummings BB, Einson J, Sousa C, Hoffman P, et al. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. Science. 2019;366(6463):351‑6. DOI: 10.1126/science.aay0256
59. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nat Commun. 2016;7(12817):6. DOI: 10.1038/ncomms12817
60. Hoang TT, Goldmuntz E, Roberts AE, Chung WK, Kline JK, Deanfield JE, et al. The Congenital Heart Disease Genetic Network Study: Cohort description. PLoS ONE. 2018;13(1):14. DOI: 10.1371/journal.pone.0191319
61. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. American Association for the Advancement of Science; 2020;369(6509):1318‑30. DOI: 10.1126/science.aaz1776
62. Muehlschlegel JD, Christodoulou DC, McKean D, Gorham J, Mazaika E, Heydarpour M, et al. Using Next-generation RNA Sequencing to Examine Ischemic Changes Induced by Cold Blood Cardioplegia on the Human Left Ventricular Myocardium Transcriptome. Anesthesiology. 2015;122(3):537‑50. DOI: 10.1097/ALN.0000000000000582
63. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078‑9. DOI: 10.1093/bioinformatics/btp352
64. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105‑11. DOI: 10.1093/bioinformatics/btp120
65. The Broad Institute. Picard Tools - By Broad Institute [En ligne]. Broad Institute [cité le 1 octobre 2020]. Disponible: https://broadinstitute.github.io/picard/
66. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841‑2. DOI: 10.1093/bioinformatics/btq033
67. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15‑21. DOI: 10.1093/bioinformatics/bts635
68. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015;12(11):1061‑3. DOI: 10.1038/nmeth.3582
69. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009;25(24):3207‑12. DOI: 10.1093/bioinformatics/btp579
70. Dobin A. STAR manual 2.7.0a. 2019.
71. Castel SE, Aguet F, Mohammadi P, GTEx Consortium, Ardlie KG, Lappalainen T. A vast resource of allelic expression data spanning human tissues. 2019; DOI: 10.1101/792911
72. McKean DM, Homsy J, Wakimoto H, Patel N, Gorham J, DePalma SR, et al. Loss of RNA expression and allele-specific expression associated with congenital heart disease. Nat Commun. 2016;7(1):12824. DOI: 10.1038/ncomms12824
73. 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;20:7. DOI: 10.1101/gr.107524.110
74. Richter F, Morton SU, Kim SW, Kitaygorodsky A, Wasson LK, Chen KM, et al. Genomic analyses implicate noncoding de novo variants in congenital heart disease. Nat Genet. 2020;52(8):769‑77. DOI: 10.1038/s41588-020-0652-z
75. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26(5):589‑95. DOI: 10.1093/bioinformatics/btp698
76. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30(7):1006‑7. DOI: 10.1093/bioinformatics/btt730

77. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81(3):559‑75. DOI: 10.1086/519795
78. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68‑74. DOI: 10.1038/nature15393
79. Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019;10(1):5436. DOI: 10.1038/s41467-019-13225-y
80. Fort A, Panousis NI, Garieri M, Antonarakis SE, Lappalainen T, Dermitzakis ET, et al. MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. Bioinforma Oxf Engl. 2017;33(12):1895‑7. DOI: 10.1093/bioinformatics/btx074
81. GENCODE. Human Release 26 (GRCh38.p10) [En ligne]. 2016. GENCODE - Human Release 26 [cité le 1 octobre 2020]. Disponible: https://www.gencodegenes.org/human/release_26.html
82. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med. Nature Publishing Group; 2017;19(2):209‑14. DOI: 10.1038/gim.2016.88
83. Liu P, Meng L, Normand EA, Xia F, Song X, Ghazi A, et al. Reanalysis of Clinical Exome Sequencing Data. N Engl J Med. 2019;380(25):2478‑80. DOI: 10.1056/NEJMc1812033
84. Ewans LJ, Schofield D, Shrestha R, Zhu Y, Gayevskiy V, Ying K, et al. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. Genet Med. Nature Publishing Group; 2018;20(12):1564‑74. DOI: 10.1038/gim.2018.39
85. Wright CF, McRae JF, Clayton S, Gallone G, Aitken S, FitzGerald TW, et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. Genet Med. Nature Publishing Group; 2018;20(10):1216‑23. DOI: 10.1038/gim.2017.246
86. Costain G, Jobling R, Walker S, Reuter MS, Snell M, Bowdin S, et al. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. Eur J Hum Genet. 2018;26(5):740‑4. DOI: 10.1038/s41431-018-0114-6
87. McLaren W. The Ensembl Variant Effect Predictor. 2016;14.
88. Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894‑9. DOI: 10.1002/humu.21517
89. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37(3):235‑41. DOI: 10.1002/humu.22932
90. Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat Commun. Nature Publishing Group; 2020;11(1):2539. DOI: 10.1038/s41467-019-12438-5
91. The GTEx Consortium. The GTEx Project [En ligne]. The GTEx ProjectDisponible: https://www.gtexportal.org/home/documentationPage#staticTextAnalysisMethods
92. Jelsig AM, Urban Z, Hucthagowder V, Nissen H, Ousager LB. Novel ELN mutation in a family with supravalvular aortic stenosis and intracranial aneurysm. Eur J Med Genet. 2017;60(2):110‑3. DOI: 10.1016/j.ejmg.2016.11.004
93. Tassabehji M, Urban Z. Congenital heart disease: Molecular diagnostics of supravalvular aortic stenosis. Methods Mol Med. 2006;126:129‑56. DOI: 10.1385/1-59745-088-X:129
94. Tan HL, Glen E, Töpf A, Hall D, O'Sullivan JJ, Sneddon L, et al. Nonsynonymous variants in the SMAD6 gene predispose to congenital cardiovascular malformation. Hum Mutat. 2012;33(4):720‑7. DOI: 10.1002/humu.22030
95. Kloth K, Bierhals T, Johannsen J, Harms FL, Juusola J, Johnson MC, et al. Biallelic variants in SMAD6 are associated with a complex cardiovascular phenotype. Hum

Genet. 2019;138(6):625‑34. DOI: 10.1007/s00439-019-02011-x
96. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat Genet. Nature Publishing Group; 2017;49(11):1593‑601. DOI: 10.1038/ng.3970
97. Caparrós-Pérez E, Teruel-Montoya R, López-Andreo MJ, Llanos MC, Rivera J, Palma-Barqueros V, et al. Comprehensive comparison of neonate and adult human platelet transcriptomes. Xue Y, directeur. PLOS ONE. 2017;12(8):e0183042. DOI: 10.1371/journal.pone.0183042
98. Vermeulen Z, Mateiu L, Dugaucquier L, Keulenaer GWD, Segers VF. Cardiac endothelial cell transcriptome in neonatal, adult, and remodeling hearts. Physiol Genomics. 2019;51(6):26. DOI: 10.1152/physiolgenomics.00002.2019
99. RefSeq. RBFOX2 RNA binding fox-1 homolog 2 [En ligne]. 2020. RBFOX2 RNA binding fox-1 homolog 2 [cité le 30 octobre 2020]. Disponible: https://www.ncbi.nlm.nih.gov/gene/23543
100. Gallagher TL, Arribere JA, Geurts PA, Exner CRT, McDonald KL, Dill KK, et al. Rbfox-regulated alternative splicing is critical for zebrafish cardiac and skeletal muscle functions. Dev Biol. 2011;359:11. DOI: 10.1016/j.ydbio.2011.08.025
101. Braeutigam C, Rago L, Rolke A, Waldmeier L, Christofori G, Winter J. The RNA-binding protein Rbfox2: an essential regulator of EMT-driven alternative splicing and a mediator of cellular invasion. Oncogene. 2014;33(9):1082‑92. DOI: 10.1038/onc.2013.50
102. Hickey EJ, Caldarone CA, McCrindle BW. Left Ventricular Hypoplasia: A Spectrum of Disease Involving the Left Ventricular Outflow Tract, Aortic Valve, and Aorta. J Am Coll Cardiol. 2012;59(1, Supplement):S43‑54. DOI: 10.1016/j.jacc.2011.04.046
103. Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science. American Association for the Advancement of Science; 2015;350(6265):1262‑6. DOI: 10.1126/science.aac9396
104. Verma SK, Deshmukh V, Nutter CA, Jaworski E, Jin W, Wadhwa L, et al. Rbfox2 function in RNA metabolism is impaired in hypoplastic left heart syndrome patient hearts. Sci Rep. Nature Publishing Group; 2016;6(1):30896. DOI: 10.1038/srep30896
105. Glessner JT, Bick AG, Ito K, Homsy JG, Rodriguez-Murillo L, Fromer M, et al. Increased Frequency of De Novo Copy Number Variants in Congenital Heart Disease by Integrative Analysis of Single Nucleotide Polymorphism Array and Exome Sequence Data. Circ Res. 2014;115:13. DOI: 10.1161/CIRCRESAHA.115.304458
106. Jin Y, Muhl L, Burmakin M, Wang Y, Duchez A-C, Betsholtz C, et al. Endoglin prevents vascular malformation by regulating flow-induced cell migration and specification through VEGFR2 signalling. Nat Cell Biol. Nature Publishing Group; 2017;19(6):639‑52. DOI: 10.1038/ncb3534
107. Sugden WW, Meissner R, Aegerter-Wilmsen T, Tsaryk R, Leonard EV, Bussmann J, et al. Endoglin controls blood vessel diameter through endothelial cell shape changes in response to haemodynamic cues. Nat Cell Biol. Nature Publishing Group; 2017;19(6):653‑65. DOI: 10.1038/ncb3528
108. Arthur HM, Ure J, Smith AJH, Renforth G, Wilson DI, Torsney E, et al. Endoglin, an Ancillary TGFβ Receptor, Is Required for Extraembryonic Angiogenesis and Plays a Key Role in Heart Development. Dev Biol. 2000;217(1):42‑53. DOI: 10.1006/dbio.1999.9534
109. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. Genet Med Off J Am Coll Med Genet. 2013;15(7):565‑74. DOI: 10.1038/gim.2013.73
110. ACMG Board of Directors. ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. Genet Med. Nature Publishing Group; 2015;17(1):68‑9. DOI: 10.1038/gim.2014.151
111. Gripp KW, Lin AE, Nicholson L, Allen W, Cramer A, Jones KL, et al. Further delineation of the phenotype resulting from *BRAF* or *MEK1* germline mutations helps

differentiate cardio-facio-cutaneous syndrome from Costello syndrome. Am J Med Genet A. 2007;143A(13):1472‑80. DOI: 10.1002/ajmg.a.31815
112. Schulz A, Albrecht B, Arici C, Van Der Burgt I, Buske A, Gillessen-Kaesbach G, et al. Mutation and phenotypic spectrum in patients with cardio-facio-cutaneous and Costello syndrome. Clin Genet. 2007;73(1):62‑70. DOI: 10.1111/j.1399-0004.2007.00931.x
113. Timal S, Hoischen A, Lehle L, Adamowicz M, Huijben K, Sykut-Cegielska J, et al. Gene identification in the congenital disorders of glycosylation type I by whole-exome sequencing. Hum Mol Genet. 2012;21(19):4151‑61. DOI: 10.1093/hmg/dds123
114. Fernlund E, Andersson O, Ellegård R, Årstrand HK, Green H, Olsson H, et al. The congenital disorder of glycosylation in PGM1 (PGM1-CDG) can cause severe cardiomyopathy and unexpected sudden cardiac death in childhood. Forensic Sci Int Genet. 2019;43:102111. DOI: 10.1016/j.fsigen.2019.06.012
115. Tegtmeyer LC, Rust S, van Scherpenzeel M, Ng BG, Losfeld M-E, Timal S, et al. Multiple Phenotypes in Phosphoglucomutase 1 Deficiency. N Engl J Med. 2014;370(6):533‑42. DOI: 10.1056/NEJMoa1206605
116. Echeverría PC, Briand P-A, Picard D. A Remodeled Hsp90 Molecular Chaperone Ensemble with the Novel Cochaperone Aarsd1 Is Required for Muscle Differentiation. Mol Cell Biol. 2016;36(8):1310‑21. DOI: 10.1128/MCB.01099-15
117. Ghosh A, Dai Y, Biswas P, Stuehr DJ. Myoglobin maturation is driven by the hsp90 chaperone machinery and by soluble guanylyl cyclase. FASEB J. 2019;33(9):9885‑96. DOI: 10.1096/fj.201802793RR

7. Supplementary data

Supplementary Tables 1 and 2: Ratios of the number of genes tested and significant when using phASER pipeline vs ASEReadCounter pipeline (Table 1) and ratios of the number of genes tested and significant comparing WGS vs WES (Table 2).

Supplementary table 1: Ratios of the number of genes tested and significant when using phASER pipelines vs ASEReadCounter pipelines

Subgroup	WES_or_WGS	Tissue	nb_genes_tested	nb_genes_sig	Subgroup	WES_or_WGS	Tissue	nb_genes_tested	nb_genes_sig	Ratio [nb_genes_tested] phASER / [nb_genes_tested] ASEReadCounter	Ratio ([nb_genes_sig] phASER - [nb_genes_sig] ASEReadCounter) / [nb_genes_sig] ASEReadCounter
ASEReadCounter_wasp	WES	ARTAORT	6122	506	phaser_wasp	WES	ARTAORT	8744	150	1.43	0.70
ASEReadCounter_wasp	WES	HRTAA	5872	740	phaser_wasp	WES	HRTAA	9202	308	1.57	0.58
ASEReadCounter_wasp	WES	HRTLV	5214	603	phaser_wasp	WES	HRTLV	8582	230	1.65	0.62
ASEReadCounter_wasp	WGS	ARTAORT	7301	826	phaser_wasp	WGS	ARTAORT	12341	174	1.69	0.79
ASEReadCounter_wasp	WGS	HRTAA	6633	1409	phaser_wasp	WGS	HRTAA	12472	409	1.88	0.71
ASEReadCounter_wasp	WGS	HRTLV	5967	1199	phaser_wasp	WGS	HRTLV	11925	336	2.00	0.72
									Average	**1.70**	**0.69**

Supplementary table 2: Ratios of the number of genes tested and significant comparing WGS vs WES

Subgroup	WES_or_WGS	Tissue	nb_genes_tested	nb_genes_sig	Subgroup	WES_or_WGS	Tissue	nb_genes_tested	nb_genes_sig	Ratio [nb_genes_tested] WGS / [nb_genes_tested] WES	Ratio [nb_genes_sig] WGS / [nb_genes_sig] WES
ASEReadCounter_wasp	WES	ARTAORT	6122	506	ASEReadCounter_wasp	WGS	ARTAORT	7301	826	1.19	1.63
ASEReadCounter_wasp	WES	HRTAA	5872	740	ASEReadCounter_wasp	WGS	HRTAA	6633	1409	1.13	1.90
ASEReadCounter_wasp	WES	HRTLV	5214	603	ASEReadCounter_wasp	WGS	HRTLV	5967	1199	1.14	1.99
									Average	**1.16**	**1.84**
phaser_wasp	WES	ARTAORT	8744	150	phaser_wasp	WGS	ARTAORT	12341	174	1.41	1.16
phaser_wasp	WES	HRTAA	9202	308	phaser_wasp	WGS	HRTAA	12472	409	1.36	1.33
phaser_wasp	WES	HRTLV	8582	230	phaser_wasp	WGS	HRTLV	11925	336	1.39	1.46
									Average	**1.39**	**1.32**

Supplementary Table 3: Count and proportion of variants for each calculated consequence for WES and WGS.

Consequence	WES count	Proportion in WES	WGS count	Proportion in WGS
intron_variant	404898	40.48%	13465063	49.36%
downstream_gene_variant	119678	11.96%	3517125	12.89%
upstream_gene_variant	119229	11.92%	3612753	13.24%
missense_variant	104360	10.43%	92060	0.34%
synonymous_variant	71916	7.19%	60732	0.22%
non_coding_transcript_variant	51150	5.11%	5497721	20.15%
splice_region_variant	36162	3.62%	38147	0.14%
non_coding_transcript_exon_variant	25103	2.51%	497046	1.82%
3_prime_UTR_variant	18257	1.83%	198370	0.73%
frameshift_variant	14099	1.41%	11727	0.04%
NMD_transcript_variant	12743	1.27%	242088	0.89%
5_prime_UTR_variant	11190	1.12%	35640	0.13%
inframe_deletion	4376	0.44%	3580	0.01%
stop_gained	2514	0.25%	2008	0.01%
inframe_insertion	1928	0.19%	2162	0.01%
splice_acceptor_variant	859	0.09%	1901	0.01%
splice_donor_variant	725	0.07%	2163	0.01%
stop_retained_variant	291	0.03%	180	0.00%
mature_miRNA_variant	242	0.02%	352	0.00%
start_lost	192	0.02%	162	0.00%
stop_lost	136	0.01%	135	0.00%
protein_altering_variant	105	0.01%	150	0.00%
coding_sequence_variant	100	0.01%	194	0.00%
start_retained_variant	59	0.01%	40	0.00%
incomplete_terminal_codon_variant	4	0.00%	7	0.00%
Total	**1000316**	**100.00%**	**27281506**	**100.00%**

Supplementary List 1: Panel genes list

A2ML1, ABCC9, ABL1, ACADVL, ACTA1, ACTA2, ACTB, ACTC1, ACTG1, ACTN2, ACVR1, ACVR2B, ACVRL1, ADA2, ADAMTS10, ADAMTS17, AFF4, AGL, AKAP9, AKT3, ALMS1, AMMECR1, ANK2, ANKRD1, APOA1, APOA5, APOB, ARHGAP31, ARID1A, ARID1B, ARL6, ARMC4, B3GALTL, B3GAT3, B3GLCT, B4GALT7, B9D1, B9D2, BAG3, BBS1, BBS10, BBS12, BBS2, BBS4, BBS5, BBS6, BBS7, BBS9, BCOR, BMPR2, BRAF, C12ORF57, C1R, C1S, CACNA1C, CACNA1D, CACNA2D1, CACNB2, CALM1, CALM2, CALM3, CALR3, CASQ2, CAV1, CAV3, CAVIN4, CBL, CBP, CBS, CC2D2A, CCDC103, CCDC114, CCDC151, CCDC28B, CCDC39, CCDC40, CCDC65, CCN1, CCNO, CDH2, CDK13, CDK9, CEP290, CFAP298, CFAP53, CFC1, CHD4, CHD7, CHRM2, CITED2, COL3A1, COL5A1, COL5A2, COX15, CPS1, CPT1A, CPT2, CREBBP, CRELD1, CRYAB, CSRP3, CTC1, CTF1, CTNNA3, DEPDC5, DES, DHCR7, DLL4, DMD, DNAAF1, DNAAF2, DNAAF3, DNAAF4, DNAAF5, DNAH11, DNAH5, DNAH8, DNAI1, DNAI2, DNAJC11, DNAJC19, DNAL1, DOCK6, DOLK, DPP6, DRC1, DSC2, DSG2, DSP, DTNA, EFEMP2, EFTUD2, EHMT1, EIF2AK4, ELAC2, ELN, EMD, ENG, EOGT, EP300, ESCO2, EVC, EVC2, EYA4, FBN1, FBN2, FGD1, FHL1, FHL2, FKRP, FKTN, FLNA, FLNC, FMR1, FOXC1, FOXC2, FOXF1, FOXH1, FOXP1, FXN, G6PC3, GAA, GATA4, GATA5, GATA6, GATAD1, GDF1, GJA1, GJA5, GJC1, GLA, GPC3, GPD1L, GYG1, HAMP, HAND1, HAND2, HCN4, HDAC8, HFE, HJV, HNRNPK, HOXA1, HRAS, ILK, INVS, IRX4, JAG1, JPH2, JUP, KANSL1, KAT6A, KAT6B, KCNA5, KCNAB2, KCND2, KCND3, KCNE1, KCNE1L, KCNE2, KCNE3, KCNE5, KCNH2, KCNJ2, KCNJ5, KCNJ8, KCNK3, KCNQ1, KCNQ2, KCNQ3, KCNT1, KDM6A, KIF7, KMT2D, KRAS, KYNU, LAMA4, LAMP2, LDB3, LDLR, LDLRAP1, LEFTY2, LMNA, LRRC50, LZTR1, MAP2K1, MAP2K2, MCIDAS, MED12, MED13L, MEIS2, MFAP5, MIB1, MKKS, MKS1, MMP21, MRPL3, MTO1, MYBPC3, MYCN, MYH11, MYH6, MYH7, MYL2, MYL3, MYLK, MYLK2, MYO18B, MYOM1, MYOZ2, MYPN, MYRF, NAA15, NDUFAF1, NDUFB11, NEBL, NEK8, NEXN, NF1, NIPBL, NKX2-5, NKX2-6, NME8, NODAL, NONO, NOTCH1, NOTCH2, NPHP3, NPPA, NR2F2, NRAS, NSD1, NTRK3, NUP155, OFD1, PCDH19, PCSK9, PDLIM3, PIK3CA, PIK3R2, PITX2, PKD1, PKD1L1, PKP2, PLN, PLOD1, PPA2, PPP1CB, PRDM16, PRDM6, PRKAG2, PRKD1, PRKG1, PRRT2, PSEN2, PTPN11, PUF60, RAB23, RAF1, RAI1, RANGRF, RASA1, RBM10, RBM20, RECQL4, RERE, RIT1, ROR2, RPGRIP1L, RRAS, RYR2, SALL1, SALL4, SCN10A, SCN1A, SCN1B, SCN2B, SCN3B, SCN4B, SCN5A, SCN8A, SCN9A, SDHA, SEMA3E, SGCD, SHOC2, SKI, SLC22A5, SLC25A20, SLC2A1, SLC2A10, SLC40A1, SLC4A3, SLC8A1, SLMAP, SMAD2, SMAD3, SMAD4, SMAD6, SMAD9, SMARCA4, SMARCB1, SMC1A, SMC3, SNTA1, SON, SOS1, SOS2, SPAG1, SPRED1, STAG2, STRA6, SYNE1, TAB2, TAZ, TBX1, TBX20, TBX3, TBX5, TCAP, TCTN2, TFAP2B, TFR2, TGDS, TGFB2, TGFB3, TGFBR1, TGFBR2, TLL1, TMEM216, TMEM231, TMEM43, TMEM67, TMEM70, TMEM94, TMPO, TNNC1, TNNI3, TNNI3K, TNNT2, TPM1, TRDN, TRIM32, TRPM4, TTC8, TTN, TTR, TXNRD2, UBR1, VCL, WDPCP, YWHAE, ZEB2, ZFPM2, ZIC3, ZMYND10

First Name: Broly Last Name: Martin

Title : Functional genomics in a diagnostic setting: ANEVA-DOT pipeline refinement and application to the Congenital Heart Disease GEnetic NEtwork Study cohort

---

Abstract

Introduction: One recent approach to gene prioritization is to integrate functional genomic information in the form of allele-specific expression (ASE) obtained from disease-relevant tissues with genomic data. Mohammadi *et al.* introduced the ANalysis of Expression Variation (ANEVA) to quantify genetic variation in gene dosage from allelic expression data in a population and developed a framework to use these variance estimates in a dosage outlier test (ANEVA-DOT). The purpose of this work was first to improve the ANEVA-DOT pipeline's accuracy by using phASER for ASE data generation. Second, we apply this novel framework to a cohort of previously undiagnosed patients with congenital heart diseases in order to identify variants responsible for the patient's disease.

Material and methods: RNA-seq data was generated from 257 heart tissue samples obtained from surgery-related discarded tissue and whole exome or genome sequencing was performed for the 220 corresponding patients. ASE data was generated either using the Genome Analysis ToolKit (GATK) ASEReadCounter utility or phASER and subsequently analysed.

Results: phASER allows for testing of a larger number of genes per sample and a larger proportion of genes previously implicated in congenital heart diseases compared with ASEReadCounter. Besides, phASER reports a lower number of outlier genes per sample and lowers the proportion of frequent outlier genes. Outlier genes are enriched for rare and deleterious variants. Using the phASER - ANEVA-DOT pipeline showed accuracy in detecting genes with pathogenic variants and led to one potential new diagnosis and the identification of several variants marked for follow-up.

Discussion: phASER - ANEVA-DOT is more accurate than previous methods and effective for gene-prioritization. This framework can be incorporated in rare disease diagnostic pipelines to use RNA-sequencing data more effectively.

Keywords

Functional genomics, allele-specific expression, RNA-seq, congenital heart diseases, gene prioritization, phASER

NOM : Broly PRENOM : Martin

Titre de Thèse : Utilisation de la génomique fonctionnelle dans un contexte diagnostique : amélioration du pipeline ANEVA-DOT et application à une cohorte de patients atteints de pathologies cardiaques congénitales.

Résumé

Introduction : Une approche récente pour la priorisation de gènes est d'intégrer les informations génomiques fonctionnelles sous forme d'expression allélique obtenues à partir de tissus pertinents pour la pathologie étudiée avec des données génomiques. Mohammadi *et al.* ont développé l'ANalyse de la VAriation d'Expression (ANEVA) pour quantifier les variations de dosage génique à partir de données d'expression allélique au sein d'une population et ont proposé un cadre analytique permettant d'utiliser ces estimations de variation dans un test de dosage aberrant (ANEVA-DOT). L'objectif de ce travail était premièrement d'améliorer la précision du pipeline ANEVA-DOT en utilisant phASER pour la génération des données d'expression alléliques. Deuxièmement, nous avons testé ce nouveau pipeline dans une cohorte de patients atteints de cardiopathies congénitales pour lesquels aucun diagnostic n'a été proposé jusqu'ici afin d'identifier des variants responsables de leurs pathologies.

Matériel et méthodes : Des données transcriptomiques ont été générées à partir de 257 échantillons de tissus cardiaque récupérés après réparation chirurgicale de pathologies cardiaques congénitales et le séquençage de l'exome ou du génome entier a été réalisé pour les 220 patients correspondants. Les données d'expression alléliques ont été générées soit à l'aide du logiciel ASEReadCounter de la suite logicielle GATK, soit à l'aide du logiciel phASER puis analysées.

Résultats : phASER permet de tester un plus grand nombre de gènes par échantillon et une plus grande proportion de gènes déjà connus pour être responsables de pathologies cardiaques congénitales comparé à ASEReadCounter. De plus, phASER rapporte un plus faible nombre de gènes ayant une expression aberrante par échantillon et limite la proportion de gènes ayant fréquemment une expression aberrante. Les gènes ayant une expression aberrante sont enrichis de variants rares et délétères. L'utilisation du pipeline phASER - ANEVA-DOT a montré une bonne précision pour la détection de gènes portant des variants pathogènes et a permis de réaliser un nouveau diagnostic probable ainsi que d'identifier plusieurs variants nécessitant des explorations complémentaires.

Discussion : phASER - ANEVA-DOT est plus précis que les méthodes précédentes et efficace pour prioriser les gènes potentiellement impliqués en pathologie. Cet outil peut être incorporé dans les pipelines analytiques diagnostiques afin d'utiliser les données transcriptomiques plus efficacement.

Mots-clefs

Génomique fonctionnelle, expression allélique, RNA-seq, pathologies cardiaques congénitales, priorisation de gènes, phASER