Molecular-Genetic and Cytogenetic Diagnosis

Yasemin Dincer*, Julian Schulz, Sandra Wilson, Christoph Marschall, Monika Y. Cohen, Volker Mall, Hanns-Georg Klein and Sebastian H. Eck

Multiple Integration and Data Annotation Study (MIDAS): improving next-generation sequencing data analysis by genotype-phenotype correlations

https://doi.org/10.1515/labmed-2017-0072
Received June 12, 2017; accepted November 14, 2017

Abstract: Next-generation sequencing (NGS) technologies in clinical diagnostics open vast opportunities through the ability to sequence all genes simultaneously at a cost and speed that is superior to traditional sequencing approaches. On the other hand, the practical implementation of NGS in routine diagnostics involves a variety of challenges, which need to be overcome. Among these are the generation, analysis and storage of large amounts of data, strict control of sequencing performance, validation of results, interpretation of detected variants and reporting. Here, we outline the Multiple Integration and Data Annotation Study, an approach for data integration in clinical diagnostics based on genotype-phenotype correlations. MIDAS aims to accelerate NGS data analysis and to enhance the validity of the results by computer-based variant prioritization using the clinical data of the patient. In this context, we present the MIDAS case reports of one patient with intellectual disability caused by a novel de novo loss-of-function variant in the GATAD2B gene [NM_020699.3: c.1426G>T (p.Glu476*)] identified by trio whole-exome sequencing, as well as two cardiac disease patients with severe phenotype and multiple variants in genes linked to cardiac arrhythmogenic disorders analyzed with multi-gene panel sequencing. Based on the data collected in the MIDAS cohort, the MIDAS software will be tested and optimized. Moreover, the MIDAS software concept can be extended modularly to include further data resources for improved data handling and interpretation in the broad field of diagnostics.

Keywords: cardiac arrhythmogenic disorders; clinical diagnostics; genotype-phenotype correlations; GATAD2B; intellectual disability; next-generation sequencing; rare diseases; whole-exome sequencing.

Introduction

Motivation for MIDAS

The diagnostics of rare genetic disorders is still a challenge for physicians today (see also the NAMSE-Initiative of the German Government, www.namse.de). Rare disorders often manifest during infancy or early childhood, yet are difficult to discern due to phenotypic variability and masking by auxiliary symptoms. Many affected individuals and support groups, who are organized in Germany in the umbrella organization ACHSE (www.achse-online.de), state that their disorder is under-diagnosed and recognized far too late. Many patients and parents of affected children have endured a diagnostic odyssey before the correct diagnosis could be obtained. During the last years, new highly parallel DNA sequencing approaches have enabled a broader, faster and more precise way to identify genetic variants [1]. These new technologies, subsumed under the term next-generation sequencing have since achieved a level of quality sufficient to justify the application in a diagnostic setting. Today’s challenges of NGS approaches are no longer the technical feasibility or costs but the evaluation of the results and the medical interpretation. Due to tremendous increasing amount of genetic data by NGS approaches in diagnostics, the integration of phenotype and genotype data becomes more and more important for data interpretation and generation of conclusive genetic reports. Therefore, the development of more powerful computational algorithms is necessary to ensure reliable and accurate bioinformatics filtering and prioritization of...
results, which is the aim of the Multiple Integration and Data Annotation Study. MIDAS focuses on the improvement of NGS data analysis for diagnostic purposes through integration of various patient data. In this context, the familial medical history of the patient and the well-documented medical examination (standardized recording of phenotype traits), the diagnostic focus (suspected disorder, associated genes) and the completeness of issued tests (coverage, diagnostic gaps as well as the classification of identified variants (class 1–5, according to Richards et al. [2]) are of prime importance. The MIDAS software aims to accelerate the evaluation of NGS results and to enhance the validity of the results by computer-based variant prioritization based on the patient data. Phenotype and genotype data of the patients collected within MIDAS will be used to test and optimize the software.

**MIDAS patient cohort and sequencing methods**

All patients were evaluated by an expert committee of clinical geneticist and extensively educated regarding potential results and incidental findings. Written informed consent was provided by each participant or legal guardian. The study was approved by the Ethics Committee (Ethik-Kommission der Bayerischen Landesärztekammer). Before enrollment, the patients had undergone routine diagnostic testing. We enrolled patients with cardiac arrhythmogenic disorders, who were analyzed by multi-gene panel sequencing, as well as children with unexplained syndromic intellectual disability and their unaffected parents (trios), who were analyzed by whole-exome sequencing (WES). NGS enrichment and library preparation, sequencing, data analysis and evaluation of results were performed in an accredited lab following state of the art guidelines and standards. Identified variants were confirmed by Sanger sequencing. Additionally, the phenotypical features of all patients were collected in a standardized way, using the Human Phenotype Ontology (HPO, [3]). The phenotype and genotype data, which was recorded and generated during MIDAS, will be used for the development and optimization of the MIDAS software.

**MIDAS software implementation and data model**

The MIDAS software is implemented in Java (Oracle Corporation, CA, USA), the graphical user interface is realized through JavaFX. It relies on a central MySQL database to store and integrate all relevant information. The implementation follows a modular approach, which each core module performing specific tasks. The NGS module is used to incorporate the results of NGS data analysis into the database using standard variant call formats (.vcf). Additionally, the NGS module allows configuration of gene panels to restrict analysis to genes relevant for the patients’ phenotype. The phenotypic feature of each patient is entered by the HPO module. This module will allow the identification of patients with similar phenotypes and the correlation of these data with the identified genotypes. Results from additional diagnostic tests like Sanger sequencing of candidate genes, identification of insertions and deletions by MLPA or structural variations by genome-wide array comparative genomic hybridization (array CGH) analysis are integrated into the database by pipeline-specific parsers. Lastly, core patient data like name, date of birth and diagnostic history are made available by the LIMS module. This allows a comprehensive overview of all relevant data of a patient through a single software interface (Figure 1).

**Case report: novel GATAD2B loss-of-function variant causes MRD18**

As one of the MIDAS patients, we present a 10-year-old boy with a combination of clinical features including delayed motor and speech development, behavioral disorders, macrocephaly and minor morphological anomalies. He is the third child of healthy, non-consanguineous parents with an unremarkable family history. His 14-year-old sister and 12-year-old brother are unaffected.
Background: intellectual disability

Intellectual disability (ID) is defined as IQ < 70 and has a prevalence ranging from 1.5 to 2% in Caucasian populations. Severe ID (IQ < 50) has a prevalence of 0.3–0.5% [4, 5]. In many cases, ID manifests not as an isolated disorder but as part of a superordinate complex syndrome with additional co-morbidities like congenital malformations, epilepsy and maladjustments such as autism spectrum disorder. It is assumed that these clinical very heterogeneous disorders have a genetic origin in at least 50% of the cases. Routine diagnostics (karyotyping, targeted sequencing in the case of a suspected disorder) is used to identify relatively common genetic causes like Down syndrome (1 in 600 newborns) or fragile-X syndrome (approx. 1 in 2500 males). It has shown that after employing routine diagnostics, a large diagnostic gap and many undiscovered causes remain. New genetic tests like micro array technologies that are able to uncover submicroscopic copy number variants (CNVs) and their application in routine diagnostics during the last decade revealed that approximately 5–20% of isolated or syndromic ID and autism spectrum disorders may be attributed to submicroscopic CNVs [6, 7]. However, even after applying these tests, approximately 60% of cases with ID remain unsolved [8].

Whole-exome sequencing enables simultaneous sequencing of all protein-coding regions of a patient. By WES, a genetic cause can be identified in approximately 27% of the patients with severe ID [4, 9–11]. Gilissen et al. [4] re-analyzed 50 patients with severe ID and negative WES by whole-genome sequencing (WGS). For 42% of these patients, a conclusive diagnosis could be obtained by WGS. But the disadvantages of WGS are currently still the very elaborate and time-consuming analysis and high costs that come along with the very large amounts of raw data. Moreover, lacking guidelines for the analysis and handling of such data as well as the difficult interpretation of intronic and intergenic regions create an obstacle for the implementation of WGS in routine diagnostics. Therefore, WES is currently recommended as a first-tier sequencing test for children with a suspected monogenic disorder (Stark et al. [12]).

Sequencing results

Karyotyping and array CGH analysis did not reveal any pathogenic variants. WES and trio analysis of the patient and his parents identified a de novo variant NM_020699.3: c.1426G>T (p.Glu476*) in exon 9 of the GATAD2B gene. The pathogenic variant was confirmed by Sanger sequencing. It was absent in the unaffected parents and not listed in population databases of the NCBI, Exome Aggregation Consortium and Genome Aggregation Database (dbSNP, ExAC, gnomAD) or in disease databases such as the Human Gene Mutation Database (HGMD® professional release, Cardiff, UK) or Online Mendelian Inheritance in Man (OMIM®, [13]).

Medical history

The boy was born at 42 weeks after a normal conception and pregnancy. Due to a pathological cardiotocograph, forceps were used during birth. With a birthweight of 4270 g, a length of 59 cm and head circumference of 39, all his growth parameters were above the 90th centile. Ultrasonography at his third day of life and at the age of 6 weeks showed normal results, macrocephaly was recorded. During the first year of life, the boy was a calm child. At the age of 1+8/12 years, a slow stato-motoric development was diagnosed and mild cerebellar ataxia was suspected. At the age of 2+5/12 years, the diagnosis of cerebellar ataxia, delayed speech development, macrocephaly and dysphagia was confirmed. He walked at 3 years (pes planovalgus was noted). At the age of 3+9/12 years, a developmental age of 1+5/12 (fine motor skills) to 2+8/12 (speech) years was diagnosed. Oppositional behavior was subject of discussion and at the age of 4 years, he was treated as an inpatient due to a behavioral disorder. Oral and tactile hypersensitivity, open bite and increased salivary flow were reported at the age of 4+10/12 years. At the age of 5, the boy had normal height (115 cm) and weight (20.5 kg), but his head circumference was above 90th centile (56 cm) and progressed to 57.1 cm (>97th centile) by age of 6. Intellectual disability (IQ = 56) was diagnosed at 6+7/12 years. A neuropsychiatric examination at the age of 7+6/12 years showed no indication for cerebellar ataxia.

The craniofacial characteristics of the patient include a prominent forehead, high frontal hairline, triangular face, mild midface retrusion, periorbital fullness, slightly prominent upper jaw and receding chin and slightly low-set ears.

Diagnosis

GATA zinc finger domain-containing 2B (GATAD2B, location: 1q21.3, OMIM.ID: #614998) encodes p66beta, a subunit of the methyl-CpG-binding protein-1 complex (MeCP1), which is repressing gene transcription by remodeling and deacetylation of methylated nucleosomes [14, 15].
Heterozygous loss-of-function variants of \textit{GATAD2B} have recently been defined to cause a recognizable syndrome with ID called Mental Retardation, Autosomal Dominant 18 (MRD18; OMIM ID: #615074). By now, eight patients have been shown to have a severely disruptive de novo variant in the coding sequence of \textit{GATAD2B} (Figure 2; [9, 17–19]) and one patient was reported with an acceptor splice-site variant in intron seven of \textit{GATAD2B} [20]. Moreover, nine patients were reported with 1q21.3 microdeletions including the entire \textit{GATAD2B} [21, 22].

The rare autosomal dominant disorder has its onset in infancy. The phenotype of patients with MRD18 is characterized by neurological features (delayed psychomotor and speech development, ID), hypotonia, long fingers/toes and craniofacial features (Table 1). Behavioral psychiatric manifestations (such as hyperactivity, tics or easy frustration) have been described in some patients.

Taken together, by considering the patient’s phenotype, we were able to identify a novel pathogenic de novo variant in the \textit{GATAD2B} gene by WES and to diagnose the rare disease MRD18 in our patient. However, the manual inspection and evaluation of variants remaining after the filtering process (in this trio case: 210 variants) is still a time-consuming process. The MIDAS software aims to accelerate the evaluation of future WES (or WGS) data by automated variant prioritization based on the patients’ phenotype. Thereby, the validity of genetic analysis results will be enhanced and faster diagnosis in patient care will be enabled. Moreover, the identification of candidate genes in MIDAS patients also provides further clinical information about patients with rare ID syndromes and thereby contributes to the specification of the associated phenotypes. Standardized recording of phenotypic traits can thereby increase the knowledge about rare diseases.

**Case reports: two patients with severe phenotype and multiple variants in genes linked to cardiac arrhythmogenic diseases**

Here we present two cardiac patients diagnosed or suspected with cardiac arrhythmogenic disorders that were analyzed within MIDAS.

**First cardiac case: medical history**

The first patient was a 10-year-old boy, who suffered from sudden cardiac death (SCD). The boy’s uncle previously had a SCD at the age of 43. The autopsy confirmed a hypertrophic cardiomyopathy.

**Background: hypertrophic cardiomyopathy**

Hypertrophic cardiomyopathy (HCM) is the most common familial heart disease with vast genetic heterogeneity. HCM is an autosomal dominant, structural disease of

---

**Figure 2:** \textit{GATAD2B} gene structure and functional domains (CR, conserved region; CR1, MBD2- and MBD3-binding; CR2, histone tail binding; [16]), marked all single nucleotide variants (SNVs) small insertions/deletions listed in the HGMD® as well as the variant found in our patient.

**Table 1:** Craniofacial features of patients with MRD18 (OMIM ID: #615074).

<table>
<thead>
<tr>
<th>Face</th>
<th>Eyes</th>
<th>Nose</th>
<th>Mouth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad forehead</td>
<td>Strabismus</td>
<td>Tubular nose</td>
<td>Broad mouth</td>
</tr>
<tr>
<td>Short philtrum</td>
<td>Hypermetropia</td>
<td>Broad nasal bridge</td>
<td>Thin upper lip</td>
</tr>
<tr>
<td></td>
<td>Narrow palpebral fissures</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hypertelorism</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deep-set eyes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the heart muscle with a prevalence of approximately 1 in 500 in Caucasian populations [23]. The disease is usually associated with an asymmetrical increase in muscle mass of the left ventricle with the interventricular septum being involved. Characteristic changes in ECG (q wave, ST segment, p wave) are the consequences. The phenotype ranges from benign forms, to forms with reduced penetrance and malignant forms with a high risk of SCD occurring already during childhood [24]. The average life expectancy is 66 years; however, the prognosis depends on the underlying molecular cause. Penetration of left ventricular hypertrophy is 95% after 55 years. Five to seven percent of the patients carry multiple variants and may have more severe or early disease expression. More than 1000 pathogenic variants have been identified in 15 HCM-associated genes so far. Most of these genes encode cardiac sarcomeric proteins. The current routine diagnostics enable detection of pathogenic variants in approximately 60% of all HCM cases. About 90% of all pathogenic variants are located in the genes for the β-myosin heavy chain (MYH7), the myosin binding protein C (MYBPC3), troponin T (TNNT2) and troponin I (TNNI3) [25].

Sequencing results

Sequencing of HCM-associated genes in MIDAS resulted in the identification of a known pathogenic variant in the MYBPC3 gene and the identification of a variant of unknown significance (VUS) in the ACTN2 gene, both in heterozygous state (Table 2). The identified variant in the MYBPC3 gene is extensively described in patients with HCM and classified as pathogenic [26–28]. The second variant in the HCM-associated ACTN2 gene may have contributed to the severe phenotype in our patient. The variant present in our patient was previously not reported. Functional prediction, using the in silico prediction programs PolyPhen2 [30], MutationTaster [31] and SIFT [32], rated the variant as deleterious. Following the standard classification guidelines of Richards et al., this variant was classified as VUS.

Second cardiac case: medical history

The second patient is a 7-year-old boy with an extreme QTc prolongation, bizarre T waves and a biphasic U wave. The boy’s father suffered from a sudden unexpected death (SUD) at the age of 20.

Background: long QT syndrome

Long QT syndrome (LQTS), a clinically and genetically heterogeneous cardiac disease, is characterized by prolonged ventricular repolarization. Long-term ECG shows prolonged frequency-adjusted QT intervals (QTc) of 440 to >500 ms. Arrhythmia, which may lead to unresponsiveness and SCD occurs in dependence to the QTc [33]. If the disease remains untreated, the 10-year survival rate is 50%. There are two types of LQTS: the common autosomal dominant Romano-Ward syndrome (RW) and the very rare autosomal recessive Jervell and Lange-Nielson syndrome (JLN). The prevalence of LQTS among the Caucasian population is at least 1 in 2500. In approximately 75% of all clinically confirmed cases, causative variants can be detected in one of the five myocardial ion channel genes. They encode one sodium channel (SCN5A, LQT3) as well as four potassium channels responsible for repolarization: KCNQ1 (LQT1), KCNH2 (LQT2), KCNE1 (LQT5), KCNE2 (LQT6). Furthermore, in rare special forms of LQTS, pathogenic variants were found in other genes: ANK2 (LQT4), KCNJ2 (LQT7), CAV3 (LQT9), SCN4B (LQT10), KCNE3, SNTA1 (LQT12). These forms are also described with highly complex phenotypes [34]. The identification of carriers of causative variants allows early, possibly pre-symptomatic treatment. In this way, the risk of cardiac events can be

Table 2: Variants identified in our patients with arrhythmogenic diseases.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Gene</th>
<th>Variant</th>
<th>Zygosity</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MYBPC3</td>
<td>NM_000256.3:c.821+1G&gt;A (IVS7+1G&gt;A, Intron 7)</td>
<td>Heterozygous</td>
<td>Pathogenic variant</td>
</tr>
<tr>
<td>1</td>
<td>ACTN2</td>
<td>NM_001103.3:c.1379A&gt;G (p.Gln460Arg)</td>
<td>Heterozygous</td>
<td>VUS</td>
</tr>
<tr>
<td>2</td>
<td>KCNJ2</td>
<td>NM_000891.2:c.436G&gt;A (p.Gly146Ser)</td>
<td>Heterozygous</td>
<td>Pathogenic variant</td>
</tr>
<tr>
<td>2</td>
<td>KCNH2</td>
<td>NM_000238.3:c.2768C&gt;T (p.Pro923Leu)</td>
<td>Heterozygous</td>
<td>VUS</td>
</tr>
<tr>
<td>2</td>
<td>PRKAG2</td>
<td>NM_016203.3:c.298G&gt;A (p.Gly100Ser)</td>
<td>Heterozygous</td>
<td>VUS</td>
</tr>
</tbody>
</table>
Sequencing results

Sequencing results of LQTS-associated genes resulted in the detection of three rare variants in our patient (Table 2). The first variant in the KCNJ2 gene is a known pathogenic variant associated with the autosomal-dominant inherited Andersen-Tawil syndrome (ATS, also known as LQT type 7). It leads to a non-synonymous substitution of glycine to serine at amino acid position 146 and was previously described causative for ATS [35]. In addition, a rare VUS was detected in the KCNH2 gene leading to the substitution of proline to leucine in the HERG protein. This variant was not previously described, yet comparable variants in the KCNH2 gene are a frequent cause of long QT syndrome. Proline at this position is conserved in homologous mammalian proteins. Functional prediction, using the in silico tools PolyPhen2, MutationTaster and SIFT, yielded contradicting results. This variant was thus classified as VUS. A second rare VUS was detected in the PRKAG2 gene, leading to a substitution of glycine by serine at amino acid position 100. For this variant a co-segregation with ventricular pre-excitation, conduction defects and HCM was shown in a Chinese family [36]. This variant seems to lead to a decreased expression of PRKAG2. The variant has an occurrence of approximately 1:15 in Asian populations [29], so it might constitute a low penetrant risk factor.

These cases exemplify the difficulties in variant classification. It could be hypothesized that the severe phenotypes of the patients are caused by a combination of the identified variants. According to guidelines and frameworks for variant classification [2], additional variants are currently only classified as VUS. For further elucidation of potential modifier effects, it may become necessary to compare sets of phenotypic features with sets of variants in a standardized way to identify possible correlations. Therefore, these data will be stored in the course of MIDAS in a comprehensive knowledge base that can be accessed for analyzing future cases with a similar phenotype.

Discussion – MIDAS concept and outlook

NGS approaches such as multi-gene panel sequencing or WES have massively accelerated the diagnosis and characterization of rare genetic diseases. However, due to the large amount of genetic data, NGS approaches are still limited by the time-consuming human interpretation of the genetic variants. Therefore, powerful computational algorithms are necessary to improve data handling in diagnostics and to accelerate the NGS data interpretation.

The main concept of the MIDAS software is the incorporation of all available clinical data of the patient in a standardized way, so that it is applicable for the NGS data analysis. Often, patients suffering from complex genetic syndromes, such as intellectual disability combined with morphological features and/or behavioral abnormalities; have to endure a diagnostic odyssey. Starting with karyotyping and array CGH analysis, up to the sequencing of single candidate genes for suspected syndromes, it can take several years until a definite diagnosis is obtained. On the other hand, human geneticists and physicians are tasked with the evaluation of all these tests, often only made available in specialized analysis software. For the interpretation, it is necessary to combine different diagnostic results, which can be very time-consuming and costly. The MIDAS software aims to mitigate this problem by integration of different diagnostic results and linking this information directly to the phenotype of the

Figure 3: MIDAS data integration concept.
individual patient. These genotype-phenotype correlations will enable a computer-based variant prioritization and thereby accelerate the identification of disease-causing variants in NGS data analysis. For the development and optimization of the MIDAS software, phenotype and genotype data of the MIDAS cohort patients will be used.

The necessity of data integration goes well beyond genetic diagnostics. For the diagnosis of many disorders, it would be hugely beneficial to integrate and correlate genetic data with further resources like imaging (nuclear magnetic resonance scans) or biochemical tests to support or exclude certain diagnoses. The MIDAS concept can be modularly extended to include further data resources (Figure 3) and provides the opportunity to improve data handling and interpretation in diagnostics.

Acknowledgments: We thank the patients and their families for the participation in the study. We thank S. Pfalzer, V. Hasselbacher, S. Eilitz, and S. Lippert for excellent technical support and M. Ziegler for critical reading of the manuscript. A consortium consisting of the Center for Human Genetics and Laboratory Diagnostics – Dr. Klein, Dr. Rost and Colleagues, Genomatix GmbH, kbo – Kliniken des Bezirks Oberbayern and IMGM Laboratories was founded to develop the MIDAS software system for data integration in diagnostics. The MIDAS project is supported by a grant of the Bavarian Ministry of Economics Affairs and Media, Energy and Technology (Grant Number: "MED-1603-0011").

References

32. SIFT. http://sift.jcvi.org/.