Special Issue Open Access

Peter M. Krawitz*

Challenges ahead for matchmaking

DOI 10.1515/itit-2016-0012
Received February 21, 2016; accepted March 11, 2016

Abstract: With every additional individual whose genome is sequenced thousands of novel variants enter the scene. It is these variants of unknown clinical significance, VUCS, that represent a great challenge to geneticists, who are dealing with high-throughput sequencing data sets. Especially in diagnostics of patients with unknown monogenic disease the joint effort of geneticists is required to find new disease gene associations. For this purpose, online platforms for matchmaking have been developed that allow clinician scientists to collaborate worldwide and to share medically relevant data. However, for a success of these tools, skills in deep phenotyping as well as new statistical approaches will be required.

Keywords: Variants of unknown clinical significance, VUCS, matchmaking, deep phenotyping, rare variant association studies.

ACM CCS: Applied computing → Life and medical sciences → Computational biology, Information systems → Information retrieval

1 Introduction

Over the recent few years, geneticists worldwide could identify hundreds of new disease genes for rare Mendelian disorders thanks to high-throughput sequencing technology. The basis of this success story is not only a very good draft of the human genome that is available now for about 15 years. Of similar importance are allele frequency data that are crucial for filtering sequence variants of an individual's genome. The 1000 genomes project, 1 KGP, generated a global reference for human genetic variation and added about 80 million single nucleotide variants to the databases of known polymorphisms [1]. A single individual differs at about 3 million positions from the haploid reference genome, that has a size of about $3.2 \cdot 10^9$ bp. When all individuals of the 1 KGP are used for filtering and only variants are considered that haven't been seen in any other sample, this number reduces to about 10^4. These rare variants are also often referred to as singletons and they are the starting point for the analysis of rare monogenic disorders. Such diseases are either caused by a single mutation that affects one copy of a gene – these disorders follow a dominant pattern of inheritance. Or, in a recessive disorder, all available copies of the gene harbor a pathogenic mutation. The incidence of a disorder may serve as a rough upper bound for the allele frequency of a pathogenic mutation. In a dominant disorder that affects less than one in a million people, the allele frequency of the disease-causing mutation is expected to be below this frequency. In a recessive disorder of this incidence, where two of such alleles must occur, the frequency cutoff is about the square root, thus about 0.001. This simple rule of thumb assumes high penetrance for the pathogenic mutations, that is, the presence of one of such alleles in a dominant disorder, or two for a recessive disorder, will cause the disease.

If you were lucky enough to find presumably pathogenic mutations in a novel gene in three or more unrelated patients with the same phenotype, you got into review in a high-ranking journal. From a statistical point of view these analyses were rather straightforward: The probability that pathogenic mutations occur by mere chance in a cohort of phenotypically similar patients is regarded so low that this null hypothesis is commonly rejected. However, this approach meets its limit when:

1. the phenotypes are extremely rare and the case groups of a single research group are too small,
2. the phenotype caused by a pathogenic allele is highly variable and penetrance is reduced, and
3. the disease causing mutations are non-coding and the allocation to a gene becomes ambiguous.

In this paper we will discuss what it will take to meet these new challenges in the identification of pathogenic alleles.

2 Networks of collaboration

The problem of small case groups can only be overcome, when clinician scientists join their efforts and share pa-

*Corresponding author: Peter M. Krawitz, Institut für Medizinische Genetik und Humangenetik, Charité – Universitätsmedizin Berlin, Berlin, Germany, e-mail: peter.krawitz@charite.de

tient data. Up-to-date this process hasn't been formalized yet, although there has been some progress on data formats. The 1 KGP helped to define a *de facto* standard for reporting sequence variants, that is called VCF [2]. This format is now used to list the variant calls of a high-throughput sequencing run and allows to compare variant data in an unambiguous way.

With phenotype data, however, it is a different story. The most effective way was to attend the annual meetings of the respective medical society, to build up one's own professional network and to make extensive use of email and telephone. However, web technology offers a lot to professionalize this scientific exchange and over the recent months many platforms for matchmaking arose. Indeed, the journal Human Mutation dedicated a whole issue to this topic and gave an overview about some of the current initiatives (Volume 36, Issue 10). The participation in any of such matchmaking platforms is surely rewarding as it is expected that more than 3000 disease-gene-associations are still to be found and the success rate is expected to grow almost exponentially with the number of contributions [3]. However, in contrast to what common sense would suggest as ideal, there is not a unique matchmaking platform, but there are plenty. For computer scientists there is a popular comic illustrating this issue for programming languages and data formats. Adapted for genetics it would probably read like this:

Situation: There are n competing platforms for matchmaking. A conversation among two geneticists: n?! That's ridiculous! We need to develop one universal platform that fits the needs of everyone! Soon: There are n+1 competing platforms.

Probably all of these platforms have a right to exist as they serve special needs and implemented features that cannot be found in the other platforms. National differences in patient data protection and sharing or the guidelines about how to classify novel variants, are likely the reason why many countries are starting their own initiatives. Here, the challenge is to find a lowest common denominator about what is allowed to exchange on an international level so that these portals can interact.

A second hurdle is that each of these platforms needs to become an independent trustee that is respected by clinician scientists that work in a highly competitive field. The VarWatch project is such an attempt that just started in Germany and that intends to act as a matchmaker for scientists with variants of unclear clinical significance, VUCS. The core idea is a "give-and-take" code of conduct. A query to VarWatch presupposes that a scientist identified some promising candidate mutations in a patient with a Mendelian disease that are yet inconclusive and he is looking for "the second patient of its kind" to clarify their status. By submitting a query, the user agrees that a collaboration is started to finalize the assessment of a variant as soon as a match was made and that more detailed phenotypic information of the patients will be exchanged.

Like all matchmaking platforms also VarWatch has to encourage contributions to obtain a critical mass to be attractive to users. Recently, VarWatch announced a collaboration with HGMD professional to ensure the quality of each query. At best, this will point VarWatch users to scientific literature they might have missed and it will help HGMD to rerate some of the mutations with false classifications, of which there are plenty.

However, a potential misconduct is also imaginable when VarWatch is used as a shortcut to get HMGD annotations. An administrator might exclude particular users with such undesired behavior, but ideally, the platform should evolve to a robust self-regulating system. A reputation system that rewards valuable microcontributions might be an option. ResearchGate, for instance, which is a fast growing professional network for scientists, encourages cooperative behavior by increasing the RG score of users that actively share their knowledge [4].

Like many resources that are important for scientific work, VarWatch is initially publicly funded, but it is required to become economically sustainable over time. It is assumed that the expertise of an independent administrator will be required beyond the funding period, if VarWatch becomes a success and is heavily used. The main task of this staff will be to support users in the process of clarifying VUCS as soon as a match was made. Currently, VarWatch is working together with several commercial software providers such as GeneTalk, JSI-Medical Systems, and Genomatix that will provide access to VarWatch integrated into their solutions. In the long run royalties from these commercial entities might finance VarWatch as efficient matchmaking adds value to their products.

Another urgent question is how to gain support of commercial diagnostics providers. It is obvious that large entities such as e. g. Myriad or Centogene, could mine their sequencing data and monetarize it as paid content (see e. g. CentoMD). However, often health insurances cover the costs of the initial diagnostics test, on which this clinical knowledge is based on. From a mere economical point of view the insurance agencies should insist that the clinical relevant data is shared and feed into free access data bases, as this will lower health care costs. If a voluntary commitment of commercial diagnostics providers turns out to be not feasible, legislators should become active.

3 Site frequency spectrum

As already mentioned in the introduction, allele frequency distributions are the key in analyzing patient's genomes. About fifty years ago Kimura developed the neutral theory of molecular evolution, that describes the allele frequency distribution, $F(x)$, for sequence variants that do not have a functional impact in a population that is constant in size and whose individuals are randomly mating [5]. For small allele frequencies, x, there is a $1/x$ dependency, that is, most alleles in the population are extremely rare. Furthermore, especially the composition of rare variants is highly population-specific. The results of 1070 recently sequenced healthy Japanese illustrate that: Of 21 million single nucleotide variants, SNVS, that were detected, more than half were novel [6]. Thus, further large sequencing projects are needed to provide suitable site frequency spectra, SFS, with high resolution for patients of different ethnicities.

4 Phenotypic overlap

For most Mendelian disorders there doesn't exist a singular pathognomonic finding, but it is rather a combination of characteristic features that guides to the clinical diagnosis. The prerequisite for a comparison of a set of symptoms is a computer searchable terminology and the Human Phenotype Ontology, HPO, has become the *de facto* standard for deep phenotyping [7]. In a tour de force Robinson et al. derived the information content of each phenotypic feature of the HPO: They annotated all known genetic disorders and computed how often a term or a descendent of it was used as a disease feature. By this means e. g. "intellectual disability" receives a lower information content as "hyperphosphatasia" as it occurs in more syndromes as a symptom.

Köhler et al. developed a tool, called Phenomizer, that allows using sets of phenotypic features to prioritize differential diagnoses [8]. In short, there might be many hundreds of diseases with intellectual disability as a feature and several others, where an elevated alkaline phosphatase occurs but the combination is highly indicative for Hyperphosphatasia with Mental Retardation Syndrome. HPMRS is a molecular pathway disease that is caused by pathogenic mutations in genes that are involved in the GPI-anchor synthesis.

For patients with dysmorphic features, image analysis technology, such as e. g. from FDNA, might also be used to support deep phenotyping. The software Face2Gene is able to predict the correct syndrome in a suggested list of ten differential diagnoses in about 80% of the cases.

Over the recent years several groups analyzed the diagnostic yield of large gene-panel sequencing approaches that were used as first line analysis for patients with suspected syndromic disorders [9, 10]. In some of these studies phenotype-based prioritization was used to rank the detected variants. Interestingly, some of the diagnoses that could be made, were beyond the phenotypic spectrum, that was known for the disease and it is therefore assumed that the phenotypic variability of many disorders is considerably higher than currently known. For inherited GPI-anchor deficiencies, IGDs, for example hyperphosphatasia was expected to be a hallmark feature. However, exome sequencing of large cohorts with patients with intellectual disability and epilepsies have also identified several novel cases with GPI-anchor deficiencies, but normal levels of alkaline phosphatase [11]. On the other hand, there were many symptoms present in these patients that fit very well to an IGD, such as muscular hypotonia, certain organ malformations, and skeletal abnormalities, but their expressivity is highly variable in this disorder.

Conceptually, frequency information can also be considered in phenotype-based prioritization algorithms. For a couple of neurological disorders Köhler et al. contributed disease annotations that incorporate whether a symptom is rare or common [12]. This quantitative approach of disease description will especially improve matching of highly variable phenotypes. Deep and quantitative phenotyping may also help to define cohorts of unsolved cases that can be further analyzed for disease gene associations. Such comprehensively characterized cohorts will also be essential for a deeper understanding of loss of function mutations with reduced penetrance [13].

5 Non-coding mutations

Let's suppose we identified the following three mutations in a cohort of patients with intellectual disability via genome sequencing:

Chr12:49.416.115G > A
Chr12:49.440.141G > C
Chr12:49.448.455C > A

A Without further knowledge it is pretty difficult to decide, whether these variants that are distributed over more than 30 kb, are disease associated. However, with the additional information that 1) all three mutations haven't been reported so far in healthy controls and 2) result in

a premature termination of the coding sequence of the gene *KMT2D* (nonsense mutations), the puzzle would be solved [14].

For non-coding mutations we not only lack theoretical models for the expected functional impact but there is also much less frequency data available. While there are almost 100.000 exomes in the public domain that can be used for filtering coding variants, we are still basically limited to about 2.500 whole genome data sets from the 1000 genomes project, when it comes to intergenic and deeply intronic variants [1]. However, both are essential when cohorts of patients with Mendelian disorders are analyzed for gene associations. In genome wide association studies, the test statistics differ for common and rare variants. Rare variants association studies, RVAS, have to work on collapsed sets of rare alleles. However, the power of these burden tests is increased only if the alleles that are subjected to the analysis, are more likely to be pathogenic [15]. In contrast, if neutral variants outweigh, it's hardly possible to detect a true disease-gene-association. For RVAS on recessive Mendelian diseases the most effective approach is a very strict allele frequency filter, prior to association testing, as most pathogenic alleles occur in less than 1 out of 10.000 healthy controls. However, in this frequency spectrum confounding from population substructure is different to common variants and the existing correction methods from GWAS cannot be applied [16]. Zhu et al. suggested a statistical framework for optimizing rare variant association studies, RVAS, on exome data for Mendelian disorders [17]. They could reproduce the results of some disease case collections that were already successfully analyzed, as e. g. the aforementioned Kabuki make-up cohort, and they could also contribute to the identification of the disease gene *TGDS* in Catel-Manzke's syndrome [18]. The probability to rank a true disease gene at the top position was higher, when cases were matched with controls that showed a similar profile of rare variants.

New large population scale sequencing projects such as the 100.000 genomes project in the UK are the precondition to extent RVAS to rare noncoding variants. In addition, our theoretical understanding of cis-interactions has to grow as classical gene coordinates will not suffice for defining appropriate intervals for collapsing rare variants. In a landmark paper by Dixon et al. topological domains, TADs, were introduced as regions of the genome that are able to functionally interact [19]. These intervals might be used for rare variant burden test. However, with a size of 2–3 Mb these regions harbor still too many neutral rare variants that dilute a signal. Even if every patient of a large case group would carry two rare non-coding pathogenic alleles in a TAD, additional filters are required for weighting. The few studies that could identify pathogenic noncoding alleles so far used e. g. epigenomic annotations for prioritization [20]. Evolutionary conservation and binding profiles of transcription factors, microRNAs and further regulatory active RNAs can serve as input data tracks for machine learning classifiers [21]. Currently the main bottleneck for such endeavors is the small training sets: In ClinVar more than 100.000 disease-causing missense and nonsense mutations are listed, whereas only a handful of pathogenic intergenic mutations are known. Hopefully, phenotype- and mutation-based matchmaking efforts will fill that gap soon.

References

1. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012, **491**(7422):56–65.
2. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al*: The variant call format and VCFtools. *Bioinformatics* 2011, **27**(15):2156–2158.
3. Krawitz P, Buske O, Zhu N, Brudno M, Robinson PN: The Genomic Birthday Paradox: How Much Is Enough? *Human mutation* 2015, **36**(10):989–997.
4. [http://www.researchgate.net/publicprofile.RGScoreFAQ.html]
5. Crow JF, Kimura M: An introduction to population genetics theory. New York, Harper & Row, 1970.
6. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S *et al*: Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications* 2015, **6**:8018.
7. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J *et al*: The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research* 2014, **42**(Database issue):D966–974.
8. Kohler S, Schulz MH, Krawitz P, Bauer S, Dolken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN: Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* 2009, **85**(4):457–464.
9. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD *et al*: Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* 2011, **3**(65):65ra64.
10. Zemojtel T, Kohler S, Mackenroth L, Jager M, Hecht J, Krawitz P, Graul-Neumann L, Doelken S, Ehmke N, Spielmann M *et al*: Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med* 2014, **6**(252):252ra123.
11. Chiyonobu T, Inoue N, Morimoto M, Kinoshita T, Murakami Y: Glycosylphosphatidylinositol (GPI) anchor deficiency caused by mutations in PIGW is associated with West syndrome and hyperphosphatasia with mental retardation syndrome. *J Med Genet* 2014, **51**(3):203–207.

12. Kohler S, Doelken SC, Rath A, Ayme S, Robinson PN: Ontological phenotype standards for neurogenetics. *Hum Mutat* 2012, **33**(9):1333–1339.
13. Ropers HH, Wienker T: Penetrance of pathogenic mutations in haploinsufficient genes for intellectual disability and related disorders. *European Journal of Medical Genetics* 2015, **58**(12):715–718.
14. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC *et al*: Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010, **42**(9):790–793.
15. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES: Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(4):E455–464.
16. Mathieson I, McVean G: Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012, **44**(3):243–246.
17. Zhu N, Heinrich V, Dickhaus T, Hecht J, Robinson PN, Mundlos S, Kamphans T, Krawitz PM: Strategies to improve the performance of rare variant association studies by optimizing the selection of controls. *Bioinformatics* 2015.
18. Ehmke N, Caliebe A, Koenig R, Kant SG, Stark Z, Cormier-Daire V, Wieczorek D, Gillessen-Kaesbach G, Hoff K, Kawalia A *et al*: Homozygous and compound-heterozygous mutations in TGDS cause Catel-Manzke syndrome. *Am J Hum Genet* 2014, **95**(6):763–770.
19. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012, **485**(7398):376–380.
20. Weedon MN, Cebola I, Patch AM, Flanagan SE, De Franco E, Caswell R, Rodriguez-Segui SA, Shaw-Smith C, Cho CH, Lango Allen H *et al*: Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* 2014, **46**(1):61–64.
21. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014, **46**(3):310–315.

Bionotes

Dr. med. Peter Krawitz

Dr. med. Peter Krawitz
Institut für Medizinische Genetik und Humangenetik, Charité – Universitätsmedizin Berlin, Berlin, Germany
peter.krawitz@charite.de