



**HAL**  
open science

# Multimodal exploration of human genome sequencing to solve the unsolved rare diseases

Kévin Yauy

► **To cite this version:**

Kévin Yauy. Multimodal exploration of human genome sequencing to solve the unsolved rare diseases. Development Biology. Université Grenoble Alpes [2020-..], 2022. English. NNT : 2022GRALV058 . tel-03870153

**HAL Id: tel-03870153**

**<https://theses.hal.science/tel-03870153v1>**

Submitted on 24 Nov 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES**

École doctorale : CSV- Chimie et Sciences du Vivant

Spécialité : Biologie du développement - Oncogénèse

Unité de recherche : IAB : Epigenetics, Environment, Cell Plasticity, Cancer (UGA / Inserm U1209 / CNRS UMR 5309)

**Exploration multimodale du séquençage de génome humain pour résoudre l'impasse diagnostique de maladies rares**

**Multimodal exploration of human genome sequencing to solve the unsolved rare diseases**

Présentée par :

**Kévin YAUY**

Direction de thèse :

**Julien THEVENON**  
Université Grenoble Alpes

Directeur de thèse

Rapporteurs :

**Alexandre REYMOND**  
PROFESSEUR, Université de Lausanne

**Peter KRAWITZ**  
PROFESSEUR, Universitätsklinikum Bonn

Thèse soutenue publiquement le **29 septembre 2022**, devant le jury composé de :

**Julien THEVENON**  
PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER,  
Université Grenoble Alpes

Directeur de thèse

**Jean-Baptiste RIVIERE**  
PROFESSEUR ASSISTANT, McGill University

Examineur

**Laurence OLIVIER-FAIVRE**  
PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER,  
CHU Dijon

Examinatrice

**Pierre RAY**  
PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER,  
Université Grenoble Alpes

Président

**Jérôme GOVIN**  
DIRECTEUR DE RECHERCHE, INSERM délégation Auvergne-Rhône-  
Alpes

Examineur

**Damien SANLAVILLE**  
PROFESSEUR DES UNIVERSITES - PRATICIEN HOSPITALIER,  
Hospices Civils de Lyon

Examineur

**Nicolas Philippe**  
DOCTEUR EN SCIENCES, SeqOne Genomics



*À mon papa,  
qui nous a quittés  
le vendredi 25 septembre 2020.*

## Remerciements / Acknowledgments

In contrast to the rest of the manuscript, the acknowledgments section will be in French. I'll start to thank my referees ("rapporteurs" in French) that accepted to review my manuscript. Dear Alexandre Reymond and Peter Krawitz, I hope you will enjoy this reading, and I'll be happy to get your comments and suggestions on this work.

### **À mes superviseurs:**

À mon directeur de thèse, Julien Thevenon. Quelle belle et longue aventure nous avons mené ensemble ! Depuis ce premier cours du DU de Dijon jusqu'à ce jour, que de beaux projets nous avons construits et liens nous avons tissés. Merci de m'avoir fait confiance depuis le début et d'avoir toujours porté une vision à long terme pour améliorer nos pratiques en génétique médicale. J'ai hâte de relever nos prochains défis à venir!

À mon responsable scientifique, Nicolas Philippe. Quel beau pari nous avons réussi ! Cette thèse est un bel exemple de synergie académie-industrie, et j'espère qu'il n'est que le début d'une longue histoire.

À mon jury de comité de suivi de thèse. Merci David Geneviève de ton compagnonnage depuis toutes ces années et d'avoir été un beau président de jury. Merci Nicolas Chatron pour ta curiosité, tes remarques toujours pertinentes et



J'espère qu'on arrivera à travailler ensemble dans l'avenir. Merci Emmanuel Barbier d'avoir porté une vision bienveillante et extérieure à la génomique dans ce comité!

À mon jury de thèse: Mille mercis de faire partie de ce jury d'exception! À Jean-Baptiste Rivière, merci de faire ce déplacement trans-atlantique, c'est un honneur d'avoir un pionnier du NGS à ce jury. À Laurence Faivre, merci pour ton énergie à faire bouger les lignes pour les patients, tu es une source d'inspiration. Merci de m'avoir envoyé à Nijmegen, quelle expérience :). À Pierre Ray, merci de ta relecture éclairée et bienveillante de ce travail de thèse! À Jérôme Govin, merci de ta disponibilité pour remplacer Sophie Rousseaux et représenter l'IAB. À Damien Sanlaville, merci Damien d'être un exemple à suivre, de ta disponibilité pour notre communauté du DES de Génétique, et de ton soutien bienveillant !

#### **Aux contributeurs éclairées de ces travaux de thèse:**

J'aimerais adresser des remerciements tout particulier aux nombreux contributeurs de ces travaux de thèse. Vous avez été des partenaires incroyables! Merci pour ces échanges riches, ce partage de compétences, et ce compagnonnage durant ces trois belles années.

À Denis Bertrand. Je t'en dois une, et même plusieurs Denis! Merci beaucoup d'avoir été présent au quotidien et de m'avoir appris ta rigueur ainsi que ton art de faire des figures. J'ai beaucoup appris en tant que scientifique grâce à toi!

À Virginie Bernard. Virginie! Mille merci pour ta bienveillance, ta disponibilité et surtout ton énergie à toute épreuve. Je me suis senti chez moi à Grenoble, et tu y es pour beaucoup. A bientôt pour de longues conversations boulot ou pas boulot ;).

À Nicolas Duforet. What a RIDE! Merci infiniment de m'avoir montré la voie de la Data Science. Mon univers scientifique ne sera plus jamais pareil grâce à toi.

À Jérôme Audoux. C'était court mais intense comme on dit! Merci à toi d'avoir toujours réussi à rendre l'impossible possible, et d'avoir été une source d'inspiration pour cette thèse et pour ma vie de cycliste ! Au plaisir d'avoir des nouvelles de vos voyages fous :).

À Raphaël Lanos. Tu es le principal acteur de ma transformation en bioinformaticien ! Un immense merci pour ta patience et ta pédagogie. Si j'ai franchi le grand pas des dictionnaires et de Python, c'est grâce à toi ! Sans oublier les jeux de société que j'ai découvert grâce à toi ;).

À Quentin Fort. C'était un plaisir de t'avoir accueilli dans l'équipe et d'avoir vu ta progression en bioinformatique ! Tu es un acteur du succès de *Genome Alert!* ;).

À l'équipe Chissé de Grenoble. Quelle belle équipe ! Malgré la période de confinement, c'était un plaisir d'avoir partagé vos locaux et tous ces moments épiques! Quentin T, c'était un honneur d'avoir été co-thésard et colloc de bureau, j'ai découvert beaucoup de choses, et pas qu'en bioinfo! Valentin, ces compliments valent aussi pour toi sur ce trio de choc. Quentin C, avec un C pour cycliste! Merci pour ces longues discussions et de m'avoir fait changer d'avis sur les montées de col à vélo ;). Laurie, merci beaucoup pour nos partages artistiques mais aussi pour ton organisation! Laura bienvenue dans l'équipe 🙌.

À l'équipe bioinfo et data de SeqOne: Si je pensais avoir des bases en programmation avant vous, je me suis trompé! Merci à toute l'équipe de m'avoir accompagné et d'avoir grandement amélioré mes compétences de bioinformaticien

et data scientist :). Special big up à Mélanie, Abdou, Nico S, Stella et Jiri pour votre bienveillance.

À ICOSA. Merci Claire, Sara et Hamza de m'avoir accompagné vers le dépôt de deux brevets sur ces travaux de thèse!

À Benoit Simard. Merci de m'avoir fait découvrir le monde des bases de données graph de façon pédagogique et élégante!

Aux collaborateurs extérieurs: Merci aux collaborateurs du CHU de Rouen (François Lecoquierre, Stephanie Baert-Desurmont, Sophie Coutant et Gael Nicolas), d'Eurofins Biomnis (Laure Raymond, Vanna Geromel) et CERBA (Armelle Luscan, Detlef Trost, Aicha Boughalem) d'avoir mené le projet *Genome Alert!* avec nous. Merci aux collaborateurs du consortium *PhenoGenius*.

**À ceux qui ont rendu cette thèse plus enrichissante:**

Au jury du Prix Sabatier d'Espeyran de l'Académie des Sciences et Lettres de Montpellier. Merci de m'avoir fait confiance et de m'avoir honoré de ce prix, merci M. Mateu pour votre pédagogie et votre bienveillance.

À Ludovic Lecordier et l'équipe de Ma Thèse en 180 secondes, c'était une très belle expérience passée à vos côtés (c'était bizarre mais surprenamment génial même exclusivement en visio!).

À la communauté du MOOC BIG. Merci à tous d'avoir participé en tant qu'organisateur, enseignant ou apprenant à cette belle aventure collective!

À la team SeqOne. Merci pour les nombreuses soirées Peacocks, offsite, vélo, vous m'avez fait découvrir un autre univers et j'en sors plus grand grâce à vous! Spéciale dédicace à Guillaume au pluriel, Danaë, Jean-Marc, Pauline, Sacha, Dimitri, Anissa,

Eric, Charles, Sanja, Matthieu, Mayl, Sam, Rafik et Gabriel. Je ne peux pas tous vous citer mais le coeur y est!

**À la belle communauté de génétique :**

Vous êtes et faites ce pourquoi être en génétique médicale est un vrai plaisir, c'est bon de se sentir en votre compagnie.

À la belle équipe AURA, avec un spécial big up à Pauline (Puppy), Pauline (Posh), Tristan, Alicia, Laury, Alexandre, Tanguy et Bertrand.

À la belle équipe de Montpellier, avec de chaleureux bisous à la team #Génétrique, quelle Dream Team, vous savez que je vous aime!

À la belle équipe de l'Ouest ou HUGO, on s'est pas mal croisé dernièrement ;).

Aux internes et amis de la SIGF / jeunes généticiens. Toutes ces soirées et ces projets que vous menez, ça me rappelle de très bon souvenirs et cela me donne confiance dans l'avenir de notre spécialité. À très vite pour de nouvelles avancées pour la génétique de demain!

**Et pour finir avec le plus important, à ma famille et à mes amis:**

À mes parents. Malgré tous les tourments et les aléas de la vie, merci d'avoir toujours été là pour moi, à tout temps. Je vous ai toujours admiré, de par les épreuves que vous avez vécu comme par la constance et l'exemplarité dont vous avez toujours fait preuve avec vos quatre garçons et avec toutes les personnes qui vous ont côtoyés. Vous êtes mes plus grands exemples, je vous aime. Papa, tu n'es plus avec nous mais tu seras toujours là dans mon cœur, mes pensées et mes valeurs.

À mes frères. À toutes ces années passées ensemble, ces expériences et ces passions partagées. Malgré la distance, nous resterons éternellement proches. À notre tour de prendre soin de Maman, l'avenir nous appartient!

À Eli. Bro, je bénis le seigneur de t'avoir placé sur ma route. Tant de choses partagées et tant de projets à continuer et tant de moments de passions à partager. Merci de partager les bons moments comme les moments difficiles, j'ai hâte d'écrire avec toi les nouvelles pages de notre histoire. Content d'avoir découvert Elisa! Il faut vraiment m'aimer beaucoup pour accepter de lire ma thèse, un grand merci à vous ;).

À mes amis de longue date : Guillaume, Annitha, Willy, Lorianne, Léna, Stéphanie, Jenny, Adeline, Claire et Randy. Tant de moments de pure folie, tant d'histoires à raconter et à vivre. Vous êtes lumineux et je suis heureux d'avoir pu croiser votre route.

À Danyel, Matthieu et Julien. A tous ces moments des études de médecine, à tous ces restaurants, ces passages montpelliérains, à toutes ses heures passées ensemble, à tous ces liens qui perdurent malgré la distance qui nous sépare.

Aux motivés de la Science Ac' : Anne, Jeanne, Jonas, Catherine, Livio, Leila, Ulysse, Martin, Leila, Tim et François. 15 ans plus tard nous y sommes, une deuxième thèse, celle de sciences. Merci pour votre temps et d'avoir fédéré des personnes magnifiques autour de vous.

À la famille Wallart. Merci pour votre soutien et pour votre bienveillance, vous êtes les bienvenus chez nous comme vous m'avez accueilli chez vous les bras ouverts.

A tous ceux que j'ai oubliés et qui se reconnaîtront, au plaisir de vous revoir!

Et surtout à Valentine. Tu es l'être qui m'a le plus compris, tu m'as tant donné et fait grandir pendant ces 9 années. Merci d'avoir été une partenaire, une amie et une écoute hors pair, merci d'être mon guide et ma lumière.

# Table of contents

<b>Manuscript summary</b>	<b>10</b>
<b>Chapter I - an introduction to genomic medicine</b>	<b>12</b>
Notions of genetics	13
The information vehicle of the living	13
From genetic information to its expression in the human body	16
Genetics or genomics?	19
Genomic medicine for rare diseases	21
Rare diseases are not rare!	21
Next Generation Sequencing, a game changer	22
A big data challenge	25
<b>Chapter II - context and motivation</b>	<b>30</b>
Thesis motivation	30
The BIG MOOC	34
3-minutes Thesis	37
<b>Chapter III - the reinterpretation challenge</b>	<b>41</b>
The variant interpretation challenge	42
Data sharing, a key element in genomic medicine	44
Motivation	45
Manuscript	48
Supplementary Materials	60
<b>Chapter IV - the phenotyping challenge</b>	<b>74</b>
Ontologies for precision medicine	75
A constellation of resources	76
Motivation	78
Manuscript	81
Supplementary Materials	112
<b>Conclusions</b>	<b>136</b>
<b>References</b>	<b>138</b>
<b>Thesis abstract</b>	<b>145</b>
<b>Résumé de la thèse</b>	<b>148</b>

## Manuscript summary

Medical genetics is a young emerging medical specialty created in France in 1995. It aims to solve the diagnostic odyssey of patients suffering from genetic diseases and coordinate their care. According to data from the Alliance Maladies Rares, 3 million people are affected in France (i.e., one person in 20), and more than 6000 different diseases are already described.

The practice of medical genetics has recently seen significant progress with the arrival of the Next Generation Sequencing (NGS), shifting from medical genetics to genomic medicine<sup>1</sup>. In 2022, we can now sequence a human genome for \$1,000 in just a few days. In contrast, the Human Genome Project initially cost €2 billion and mobilized an international research effort over several years. The limitation is no longer the sequencing but the bioinformatic processing of the massive genomic data generated by the NGS and their clinical interpretation. The democratization of genome sequencing has made it possible to discover the molecular involvement of many new genes at the origin of pediatric and adult rare diseases<sup>2,3</sup>. Genetic tests are increasingly prescribed and included in healthcare systems due to the decreasing sequencing costs, increasing performance of technologies (cloud computing, etc.), and new applications of genomic medicine<sup>4,5</sup>. However, many patients remain undiagnosed after genome sequencing.



Using bioinformatics and data science, my thesis project aimed to manage current bottlenecks of genomic medicine in patient care to improve rare disease diagnoses. Even if I was the main contributor to the present work, it was only possible to achieve thanks to the fantastic team in SeqOne Genomics and CHU Grenoble Alpes.

This manuscript contains four chapters, starting with definitions of the notions and concepts in Genomic medicine mainly adapted from the “MOOC BiG - Introduction to Bioinformatics and Genomic Medicine” I co-led. The second chapter sets the context for this thesis and presents a report on pedagogical works I realized during this Ph.D. Then I described two main scientific projects I led during this Ph.D. The third chapter is about *Genome Alert!*, an open-source method that monthly reassesses variant pathogenicity and gene-phenotype associations by data-mining the collaborative ClinVar database while highlighting changes likely to impact diagnosis. The final chapter will narrate the development of *PhenoGenius*, a machine-learning technique to thwart fuzzy clinical descriptions from physicians' phenotyping.

# Chapter I - an introduction to genomic medicine

## *Towards a care pathway in genomic medicine*

Genomic medicine profoundly changes medical practice and allows unprecedented access to precise diagnoses, personalized care, preventive actions, and targeted therapeutic adaptations. This introduction chapter aimed to define notions in genomic medicine necessary to understand the scientific advancements presented in the following chapters. As genomic medicine relies on human genetic characteristics, I first explained the basic concepts of genetics and the diversity of genetic variations. As I was interested in improving the diagnosis of rare diseases, I presented the NGS revolution that significantly improved their diagnostic yield. This revolution has also profoundly changed medical practices and has allowed the rise of genomic and precision medicine, an evolution to which my work is linked.

# Notions of genetics

## The information vehicle of the living

*DNA is the information carrier of life*

DNA, Deoxyribonucleic acid, is a macromolecule (large molecule) central in the cell and carries the capacity to support information within it. This information and the ability to transmit it is fundamental for the living. Indeed, it is thanks to the DNA molecule that organisms living on earth can transfer the information necessary for self-maintenance of their system from one generation to the next. But also, through mechanisms of random mutation of the information, evolutionary changes rise to individuals whose biological functions or physical traits are altered. Evolution, as related to genetics and as described by Darwin <sup>6</sup>, refers to the process by which living organisms change over time through changes in the genome <sup>7</sup>.

*History of 3.2 billion letters in 23 volumes*

DNA is a macromolecule composed of an assembly of smaller molecules, nucleotides (or bases) Adenine, Thymine, Guanine, and Cytosine (A, T, G, and C) <sup>8</sup>. These nucleotides connect to form a very long DNA molecule. Two single-stranded DNA molecules stick together by weak chemical bonds, matching Thymines with Adenines and Cytosines with Guanines.

As described by Watson and Crick in 1953 based on Franklin's work, this double-stranded structure takes the form of a DNA double helix resembling a ladder

whose backbone is a sequence of nucleotides, and the ladder's steps are the bases linked two by two <sup>9,10</sup>.

In our cells, protected in the nucleus, there is not only one double-stranded molecule but 23 pairs of these molecules, which correspond to 23 pairs of chromosomes. 22 pairs of these are called autosomes and are similar for men and women. The 23rd pair corresponds to the sex-indicating chromosomes called gonosomes: a pair of X for women and the X and Y chromosome for men. We speak about pairs of chromosomes because our cells have two almost identical copies of their genetic information. We also say that our cells are diploid. If we measure the total length of human DNA, we will result in a sequence of 2 copies of 3.2 billion base pairs <sup>11</sup>.

*The genetic code enables us to go from sequences to proteins*

The DNA molecule is the information storage medium for our cells. The cell encrypts the information using specific successions of bases, like our computers using binary sequences of 0 and 1. Similar to binary, which, to make a byte, cuts the information into blocks of 8, the cell cuts the information in particular regions into blocks of 3, triplets, also named codons. These codons allow correspondence between the genetic information in the DNA and the production of proteins from amino acids. This corresponds to the genetic code deciphered by the Nirenberg team from 1961-1966 <sup>12</sup> (Figure 1).

We can determine the sequence of nucleotides via sequencing techniques. The entire sequence of nucleotides is called the 'genome', which provides the entire genetic information of an individual. We also use the word "genome" for the process of "genome sequencing".

In summary, DNA is the human body's information molecule. This information is a sequence of A, T, G, and C nucleotides in succession. For human beings, the information is present in 2 almost identical copies of about 3.2 billion base pairs. The genetic code is the "rosetta stone" between blocks of three nucleotides and amino acids (needed to build proteins). There are techniques called sequencing that allow us to know the order of nucleotides.

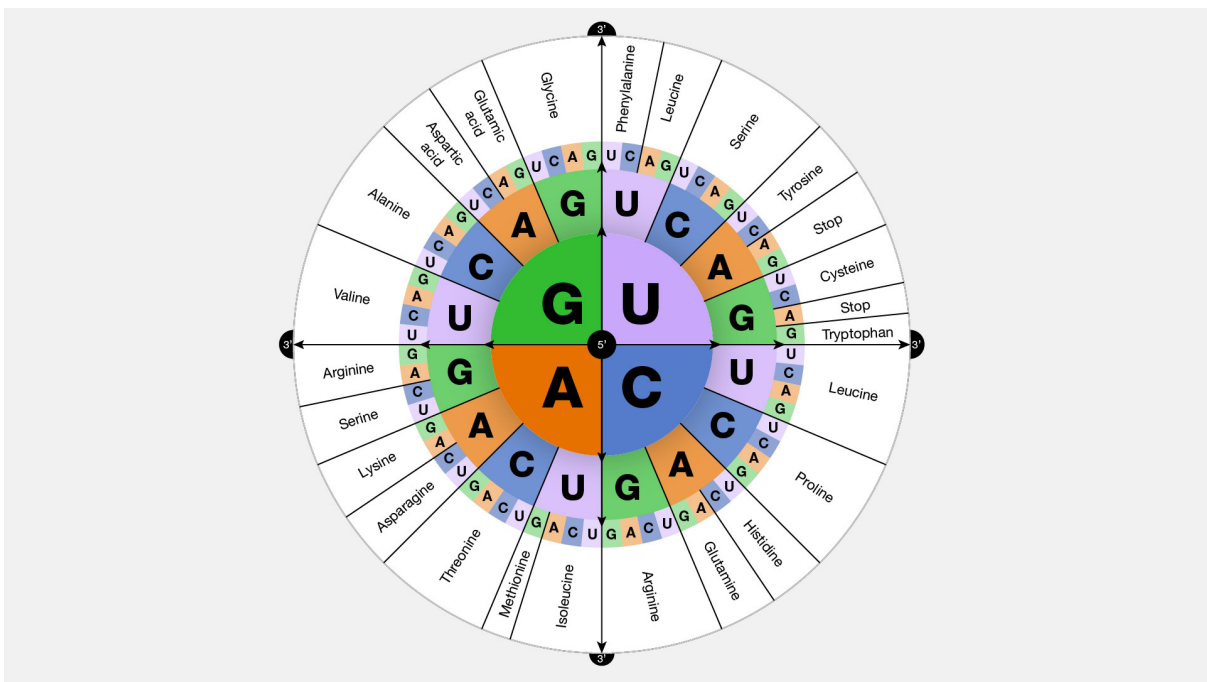


Figure 1. Genetic code representation, from NIH National Human Genome Research Institute (<https://www.genome.gov/genetics-glossary/Genetic-Code>).

## From genetic information to its expression in the human body

*Genotype and phenotype represent our genetic and physical features*

As a human, we are not only determined by our genetic heritage. We are, as an organism, the simultaneous expression of our genetic heritage influenced by the environment. Our features, our physical and psychological characteristics are what we call the phenotype. It is what is visible and recognizable about an individual. Some phenotypic features seem to be transmitted from one generation to the next and are most probably linked to genetic traits. Following the example of the word “phenotype”, we have defined the term “genotype”, which corresponds to an individual's genetic features <sup>13</sup>.

*The coding parts of our genome are the templates of our proteins*

From a genetic perspective, the word gene defines a region of DNA that can be transcribed into ribonucleic acid or RNA. Many regions in the genome can be transcribed into RNA; some of these allow the cell to produce proteins, which are biological tools in a broader sense <sup>8</sup>. These protein-producing genes are called "coding genes," and there are just over 20,000 known coding genes in human beings <sup>14</sup>. In the RNA of these coding genes, there are two types of regions, which are actually used to make a protein, called “exons”, and regions with no link with protein production, called “introns”.

To produce a protein, the RNA of a coding gene will undergo a step that eliminates the introns and keeps only the exons. This step is called splicing and is very important to obtain a messenger RNA (mRNA) which will leave the nucleus to allow the production of a protein <sup>8</sup>. Note that to determine all the genetic information contained in these coding regions, we would sequence all DNA regions corresponding to the exons of nearly 20,000 genes. This is known as exome sequencing, and in everyday language, we say we "do an exome" <sup>15</sup>.

*Each human being has their own genetic variations*

Between individuals of the same species, we are genetically very close but not identical. We are all made up of millions of genetic variations, the vast majority of which are polymorphism (no effect on the phenotype). Moreover, these genetic variations can identify an individual and the population from which he comes and shares genetic features <sup>16</sup>.

These variations can affect a base (for example, an Adenine becomes a Guanine), and we speak then about "SNV" (for Single Nucleotide Variant). When the variation affects several thousands of bases or even millions of bases, we speak of "SV" (for Structural Variation, Variation in the structure of the DNA) <sup>17</sup> (Figure 2). These can be called translocations (2 chromosomes exchanging genetic information), insertions, or inversions of genetic information. Among these SVs, there can be a gain of

genetic information (duplication, triplication, ...) or a loss of information (deletion) that we define as CNV for Copy Number Variation <sup>18</sup>.

### Structural variations

#### Inversion



#### Translocation



#### Deletion



#### Duplication



#### Insertion



Figure 2. All types of structural variations reported in the human genome. Adapted from James Hutson from Garvan Institute for Medical Research.



In a nutshell, the phenotype is a set of an individual's visible "physical" features. The genotype is a set of genetic features of an individual. There is a correlation between the genotype and the phenotype of an individual. Genes are subunits of genetic information linked to production, some of which are known as "coding" and allow the production of proteins. We are all made of thousands of variations of all sizes, often benign, but some can have substantial deleterious impacts.

## Genetics or genomics?

*Genetics is a study of genes and heredity*

Genetics is the study of heredity, or how the characteristics of living organisms are transmitted from one generation to the next via DNA. Medical genetics is the branch of medicine that involves diagnosing and managing hereditary disorders, using genetic knowledge in human diseases. Because sequencing techniques were limited in scale until the 2010s, it has typically focused on variations in a single gene when determining the cause of a health condition <sup>19</sup>.

*Genomics describes the study of the whole genome*

Genomics describes the study of the whole person's genetic information (the genome) (Figure 3). In addition to the medical genetics benefits, genomic medicine offers new possibilities such as pharmacogenomics but also provides new challenges like incidental finding management <sup>20</sup>.

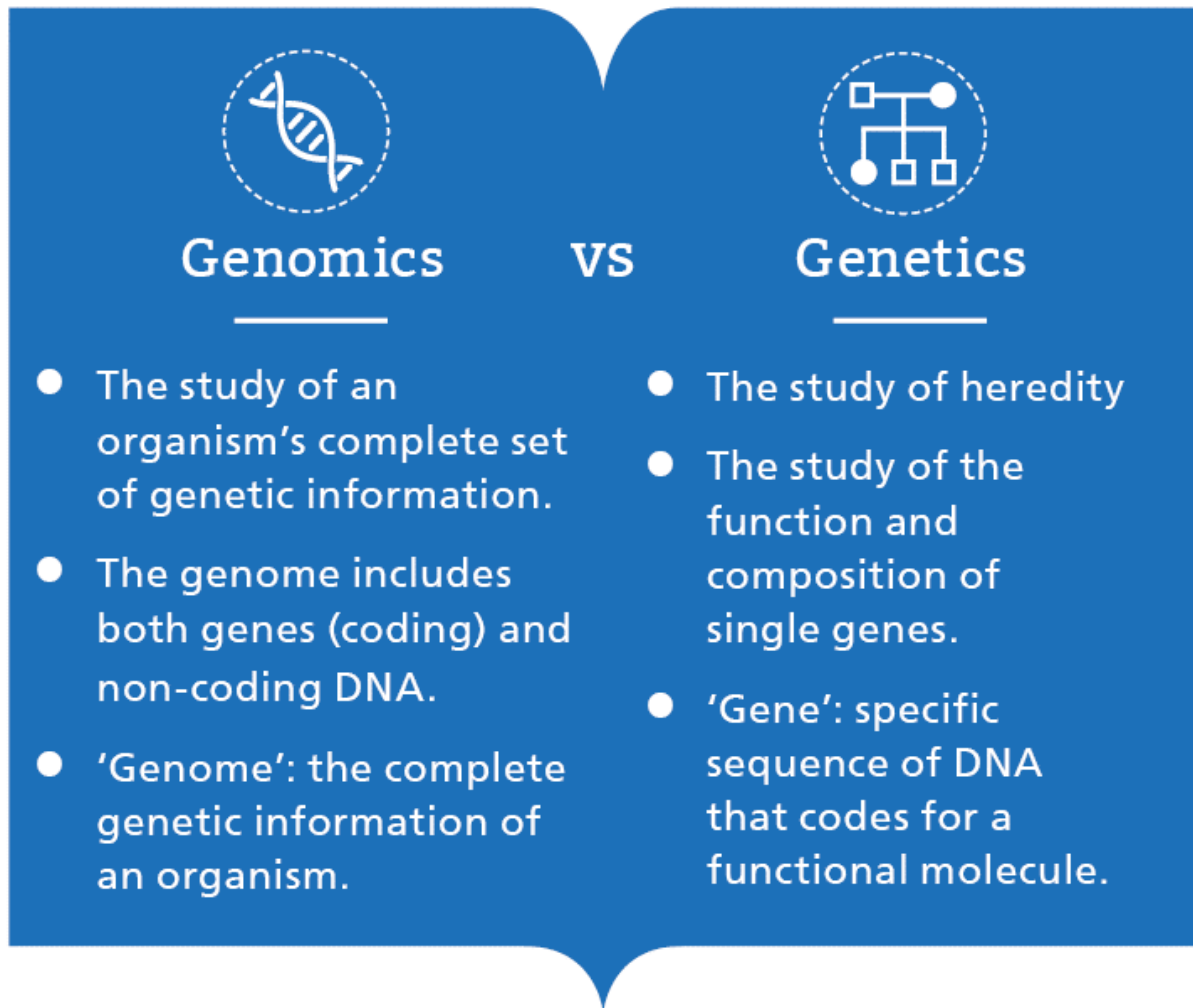


Figure 3. Genomics vs Genetics Fact Sheet. Adapted from the Genomic Education Programme from the NHS' Health Education England.

Overall, this section defined biological concepts and genetics vocabulary used in the medical interpretation of genome sequencing. In the next paragraph, I introduced applications of these concepts into rare disease diagnosis and provided an overview of the revolution of NGS in clinical practice.

# Genomic medicine for rare diseases

## Rare diseases are not rare!

According to EURORDIS - Rare Diseases Europe organization (<https://www.eurordis.org/>) statistics, a rare disease in Europe is a disease affecting less than 1 in 2000 people. 3% of births are affected, and 7-8% of adults live with a rare disease among the 6000 currently described (Figure 4). In Europe, 25,000,000 people are concerned, 50% are children under five years old, and rare diseases cause 10% of deaths in those under five years old. Although rare diseases are individually rare, they are collectively frequent. The conditions are often chronic, severe and lead to an alteration in the quality of life. 10% of people lose autonomy, and 50% of people have a motor, sensory or intellectual deficit.

An estimated 72% of rare diseases are genetic in origin. This genetic origin is essentially monogenic, i.e., the alteration of a unique gene is responsible for the rare disease. The diagnosis of rare diseases is complex due to the clinical and genetic diversity (heterogeneity) of these diseases. This diagnostic challenge is responsible for delays in diagnosis.

The diagnosis of a rare disease is essential for several reasons<sup>21</sup>, e.g.

- Personalize care: specific follow-up can be initiated according to known disease complications and offer appropriate care.

- Genetic counseling: evaluate the risk to future offspring or close relatives inheriting the condition
- Disability recognition: Informing patients and families about the disease and getting disability recognition by society.

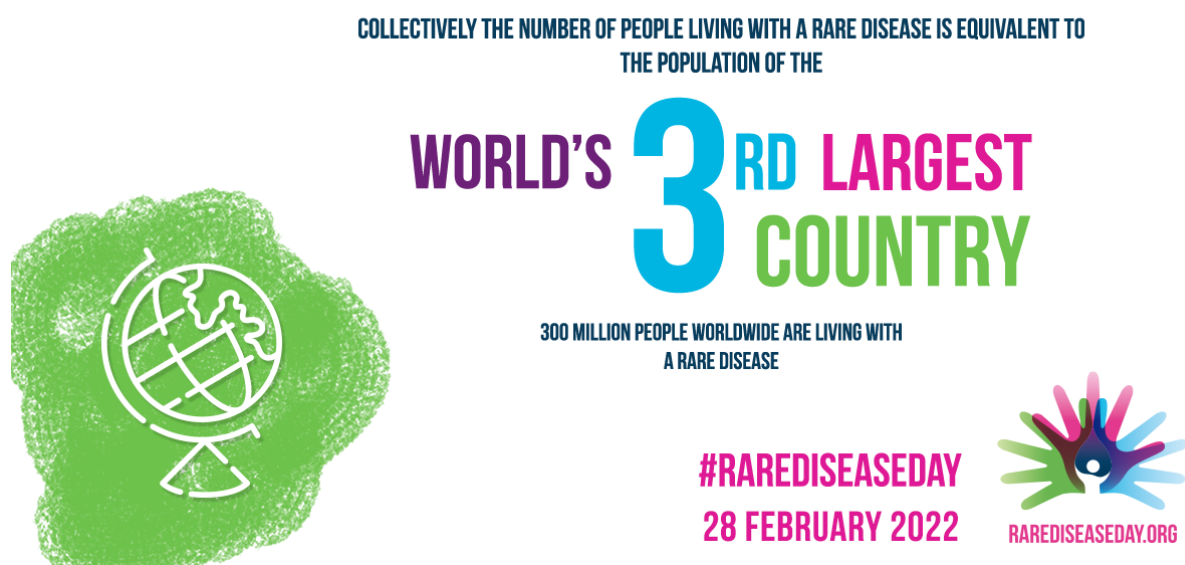


Figure 4. Infographic for rare disease day, a yearly event that raises awareness of rare diseases for the general public.

## Next Generation Sequencing, a game changer

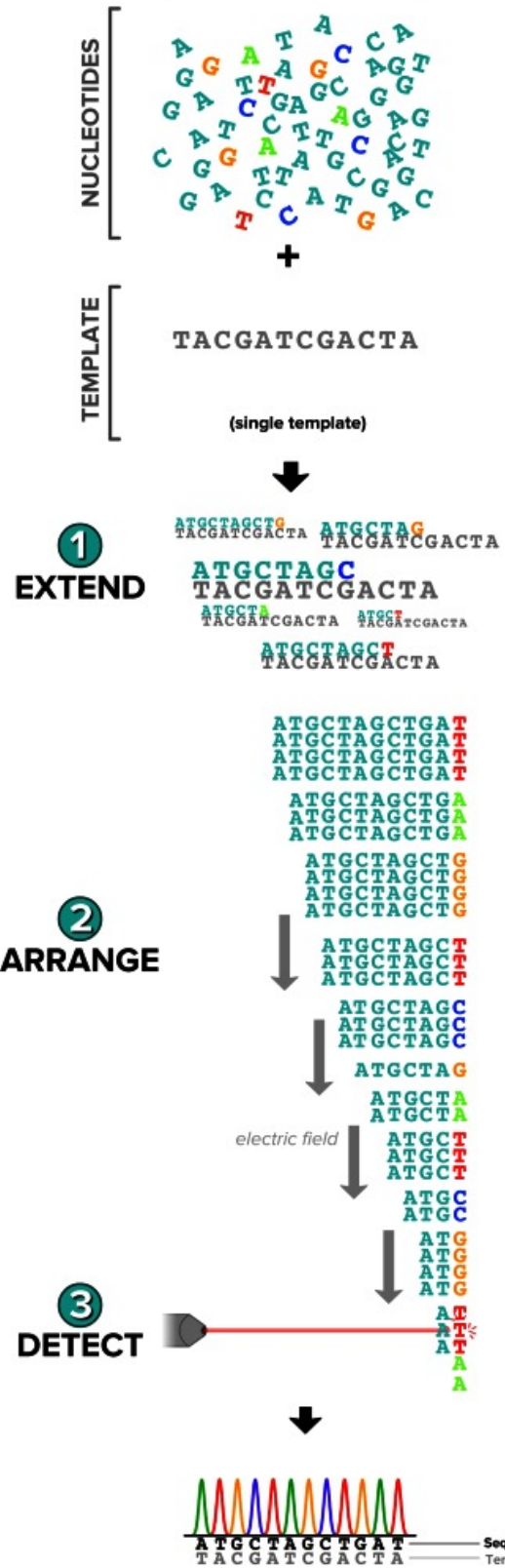
### *A massively parallel sequencing*

While Sanger sequencing, a historical technique, allows the analysis of only one DNA fragment at a time, NGS can sequence many fragments simultaneously, hence its name “massively parallel sequencing”.

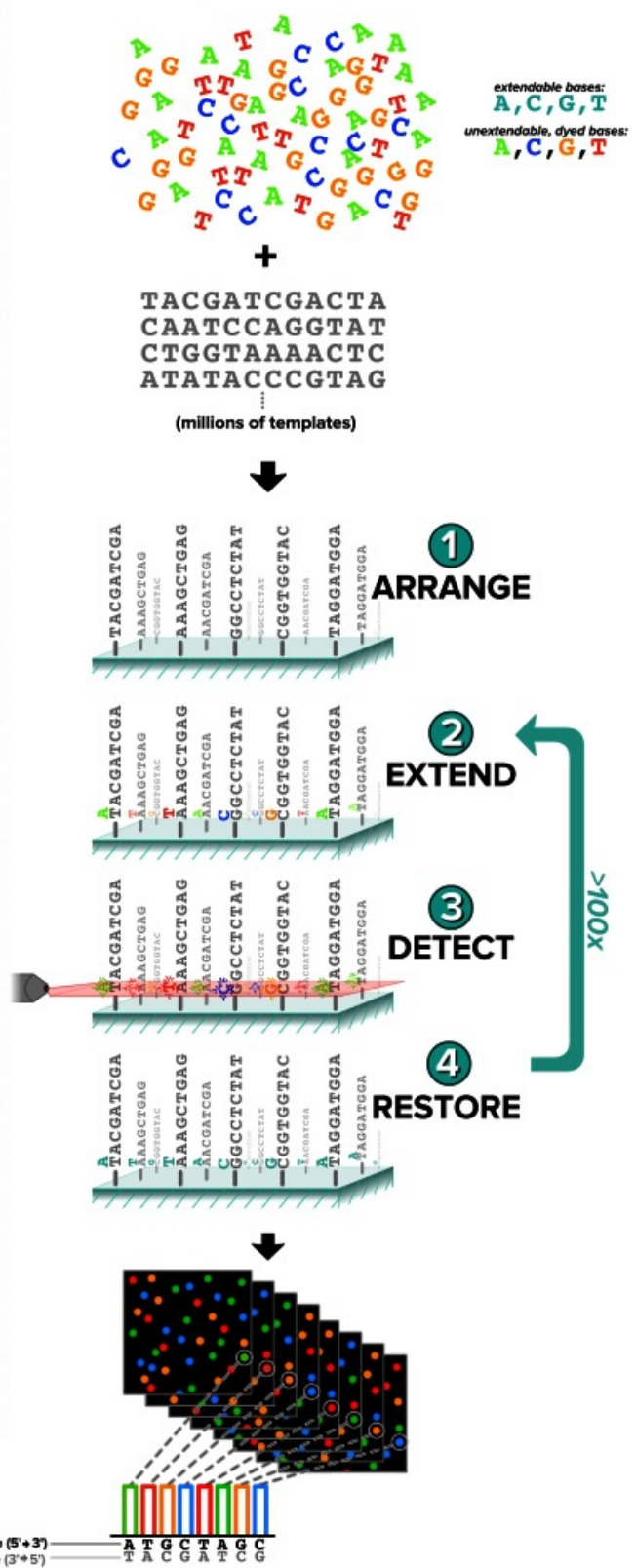
The main NGS technology available is based on the sequencing by synthesis of Polymerase Chain Reaction (PCR) colonies, developed by Shendure *et al.*<sup>22</sup> and proposed by Illumina company. In brief, the base calling or identification with Illumina sequencing technology is obtaining nucleotide sequences from fluorescent signals. DNA fragments are fixed on a plate called a “flow cell” and amplified using PCR. PCR relies on a DNA polymerase that “reads” the existing DNA strands to create two new strands that match the existing ones, thus replicating the DNA. This amplification allows to take pictures of big enough fluorescent DNA colonies (or clusters) and obtain DNA sequences (Figure 5).

In detail, sequencing takes place in several cycles, each corresponding to a base's identification. Each cycle consists of several steps. In the first step, the polymerase and the four types of nucleotides, which are fluorescent and carry a chain terminator, are in solution. The polymerase will incorporate only one nucleotide during this step because the chain terminator prevents the other nucleotides from binding. During the second step, a camera will take a picture of the flow cell to highlight the fluorescence signals corresponding to the incorporation of the nucleotide. The third step is a washing step to get rid of the fluorescence and the chain terminator. The cycles will follow one another until the complete sequence is obtained.

# SANGER:



# NGS:



extendable bases:  
A, C, G, T  
unextendable, dyed bases:  
A, C, G, T

Figure 5. Illustration of massively parallelized NGS compared to Sanger sequencing.

Adapted from Muzzet, Evans, and Lieber (2015) <sup>23</sup>

### *A shift from medical genetics to genomic medicine*

The NGS revolution has led the transition from medical genetics to the genomic medicine era: from the end of the 2010s to the present day, genetic sequencing has gone from a few genes to the whole genome <sup>24</sup>. With genome sequencing accessibility, rare disease diagnosis shifted from this phenotype-first approach to a genotype-first approach <sup>25</sup>. The possibility of exploring all human genes in NGS makes it possible to respond to the extreme heterogeneity of rare diseases. For example, in the context of intellectual disability, more than 1000 different genes are involved. Diagnostic yields for many other rare diseases have been greatly improved since its appearance <sup>26</sup>.

Still, getting the DNA sequence is insufficient to provide a patient with a genetic diagnosis. While sequencing is no longer limited, several studies have pointed out that NGS data processing constitutes a limitation called the bioinformatics bottleneck of genomic medicine <sup>27,28</sup>.

## **A big data challenge**

As a complement to clinical and laboratory genetics, bioinformatics and data science have become critical elements in genomic medicine. It responds to a very

concrete need linked to the rapid increase in the volume of tests, the volume of data, the volume of clinical knowledge, and the type of data.

#### *From sequences to variant interpretation*

The processing of raw sequencing data from NGS to obtain a list of characterized alterations and their interpretation in the medical context requires the use of various tools whose algorithm has been optimized for the following specific tasks <sup>29</sup>:

- The alignment corresponds to finding the place of each DNA read in the human genome by comparing the sequence of the read to a human reference genome built by the Genome Resource Consortium (<https://www.ncbi.nlm.nih.gov/grc>). We obtain millions of reads to compare for each genome sequencing.
- The next step is called variant calling. The general idea is to detect differences between the patient's DNA and the reference genome used for the alignment. Different algorithms are required depending on the type of alterations sought.
- Finally, the detected alterations are annotated by comparing the positions and types of events with the different available databases (frequency in population, effect of variants, etc...). This task is the most diverse and evolutive in genomic medicine.

Improving these tools allows for the detection and medical interpretation of novel variants that can lead to diagnostic solving <sup>30</sup>.



### *Rare disease knowledge overwhelms human learning abilities*

The formulation of medical diagnostic hypotheses relied on identifying symptoms and evaluating their joint associations with diseases. Although such associations could be easily done for common disorders, they pose a significant challenge for rare diseases where over 6000 diseases must be matched with clinical features. Rare disease knowledge overwhelms human learning abilities and is constantly increasing<sup>31</sup>. Moreover, in the genome sequencing era, rare disease diagnosis is currently limited by human bottlenecks such as the time-consuming clinical reassessment step, where physicians reanalyze clinical observations according to the genome sequencing analysis<sup>32</sup>.

This bottleneck is currently tackled by computational phenotype analysis development aiming to better integrate clinical data into genome analysis workflow<sup>33</sup>, a fundamental step to the rise of precision medicine.

### *Precision medicine using artificial intelligence*

Precision medicine aims to define disease at a higher resolution by genomic and other technologies to enable more precise targeting of disease subgroups to improve diagnosis, prognosis, and medical treatment<sup>34</sup>. Citing Peter M. Krawitz, “human and artificial intelligence (AI) need to join efforts” is the only way to succeed in this medicine revolution.

Based on structured clinical, biological, and imaging data of an Electronic Health Record (EHR), precision medicine has proven to benefit healthcare through

phenotypically rich EHR and large sequencing cohorts. Pilot studies such as DiscovEHR Collaboration between the Regeneron Genetics Center and Geisinger Health System reported valuable insights and redefinition of genetic diseases as hundreds of individuals with rare variants are linked to novel phenotypes <sup>35,36</sup>. Genomic England's "100,000 Genomes Project" recently reported its first insights on precision medicine's impact on rare diseases diagnostic yield <sup>37</sup>.

Moreover, deep phenotyping provides additional features to characterize patients better <sup>38</sup>. As an example, recent studies have demonstrated that facial analysis technologies may support the capabilities of expert clinicians in syndrome identification, even undescribed by clinical geneticists <sup>39</sup>. Indeed numerous genetic disorders may have recognizable facial features, accessible through expert clinical examination by an expert in the field. For non-expert clinicians, considering hundreds of diagnostic hypotheses is rarely feasible. Facial image analysis frameworks such as GestaltMatcher use computer vision and deep-learning algorithms that quantify similarities to hundreds of syndromes and identify facial phenotype descriptors <sup>40</sup> (Figure 6).

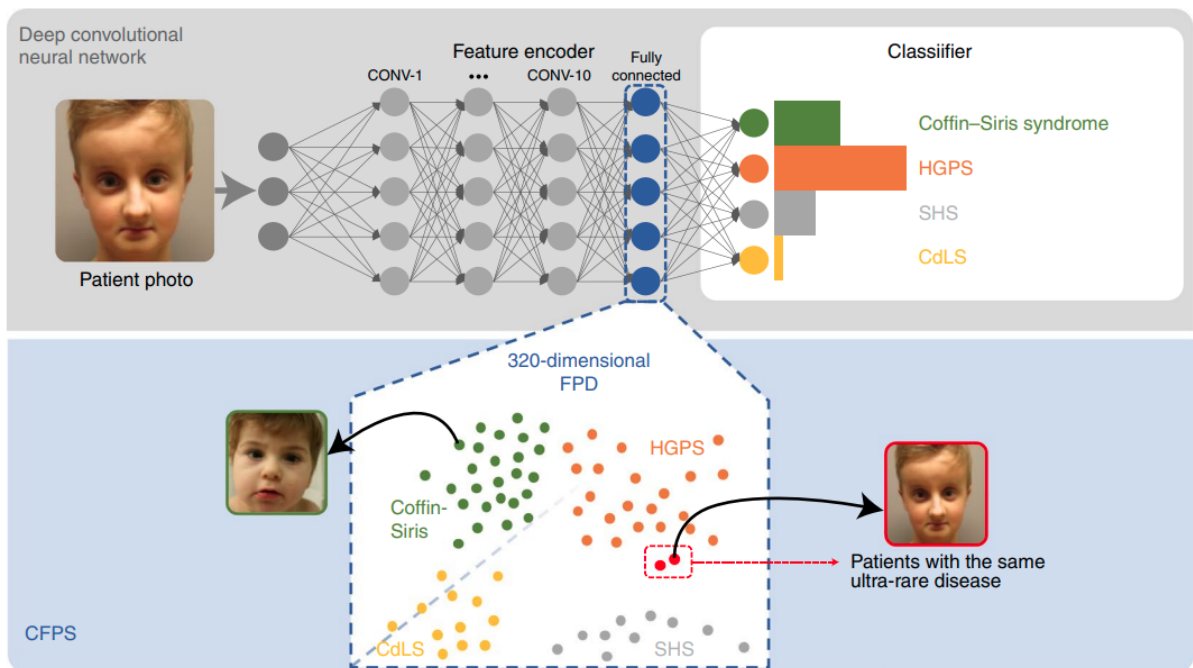


Figure 6. Illustration of GestaltMatcher. Using a deep convolutional neural network, GestaltMatcher enables clinicians to match patients with facial similarity and thus, possibly diagnose patients with an ultra-rare disorder or delineate a new syndrome in similar patients <sup>40</sup>.

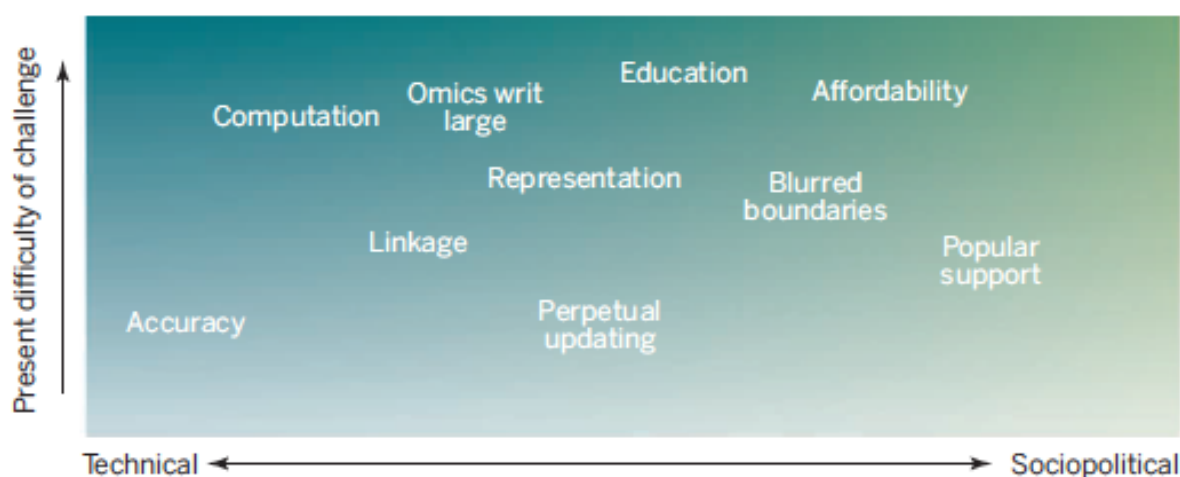
In a nutshell, rare disease diagnosis relies incrementally on machine learning and bioinformatics programs to exploit clinical and sequencing data to improve patient care.

## Chapter II - context and motivation

*It's a long way to the top if you want genomic medicine...*

### Thesis motivation

I described in Chapter I basic notions of medical genetics and introduced the increasing need for bioinformatics and data science in genomic medicine. A lot is still to be built to implement precision medicine in the routine clinic (Figure 7).



**Moving toward precision medicine.** Ten challenges for achieving precision medicine are qualitatively ordered on the x axis by how much they are intrinsically technical versus sociopolitical challenges. The y axis qualitatively orders the difficulty each challenge currently presents if we are to attain the widely articulated goals for precision medicine.

Figure 7. Ten challenges for achieving precision medicine. Adapted from Kohane Science (2015) <sup>41</sup>.

Nation-wide and population genomics programs have increased the global knowledge of the population genetic variability and provided access to genome sequencing in healthcare <sup>42,43</sup>. Despite the accessibility of genome sequencing in clinical routine, a majority of patients are still in a diagnostic deadlock <sup>37</sup>, meaning that all of the investigations currently available in clinical practice to determine the precise cause of the disease are exhausted. It concerns patients suffering from an atypical form of a known disease or a disease whose genetic or other cause has not yet been recognized. One issue is in the management of the data: we can no longer read everything, we can no longer learn to diagnose everything, yet we must follow the progress and bring accurate information to the patients. Moreover, for medical genetic practitioners, new skills are expected to meet the challenges of tomorrow. In France, the national sequencing plan “France Médecine Génomique 2025” implementation illustrates the need to acquire notions of sequencing, algorithms, data analysis, modeling, statistics, and massive data management (<https://pfm2025.aviesan.fr/le-plan/formation-continue-et-professionnelle/>).

To pursue the adoption of genomic medicine in healthcare, bioinformatics and data science concepts must be learned to understand current and future multi-omics techniques and dialogue with bioinformaticians responsible for sequencing analysis. In addition, new methods using AI need to be invented to decipher knowledge from

genome sequencing data and be accessible to the community to ensure genomic medicine implementation.

In this context, I did my Ph.D. in the framework of an academic-industrial partnership (Université Grenoble Alpes - SeqOne Genomics) with three objectives: First, providing a resource to teach notions of genomic medicine and bioinformatics with my academic team. Second, performing scientific explorations and developing methods to manage current bottlenecks in genomic medicine both with academic and industrial groups. Third, industrializing these methods to maintain them and make them accessible to the community with the industrial team.

Communication and teaching are crucial elements in raising awareness of genomic medicine and supporting necessary changes in the practice of the medical community. In the following paragraphs of this Chapter, I reported the two main actions I led: developing a Massive Online Open Course (MOOC) in bioinformatics for genomic medicine and participating in the 3-Minutes Thesis competition.

In the following chapters of the manuscript, I then described the two main focuses of my scientific work:

- Chapter III: The perpetual updating challenge and reinterpretation bottleneck of previously unsolved genomic analysis detailed. New medical discoveries could solve previously undiagnosed patients, but no clear workflow or recommendations exist to provide this crucial task to the clinic.

- Chapter IV: The clinical data computation challenge, where medical coding or physician's phenotyping are reported heterogeneous. This scanty phenotyping is a significant barrier to precision medicine, to exploit medical data, and provide computationally-detected clinically relevant groups of patients.

To be noticed, I also participated in developing bioinformatics pipelines to improve the clinical workflow of genomic analysis <sup>44,45</sup>, provide better insights into precision medicine in breast cancer <sup>46</sup> and explore applications of deep learning methods in Kabuki syndrome diagnosis <sup>39</sup>.

## The BIG MOOC

Following the rise of precision medicine and ensuring its adoption in rare disease management, I was regularly asked to teach lectures in Bioinformatics by the community. I have organized and taught Bioinformatics lectures to French residents since 2018. But faced with the high demand that bioinformatics teachers cannot keep up with, I started to build the project of a MOOC in bioinformatics for genomic medicine and developed it during my Ph.D. Joined by Evan Gouy as a co-project leader and Julien Thevenon as coordinator, this work led to the construction of the educational storytelling that inspired the sections of the Ph.D. manuscript.

The MOOC BiG "Introduction to Bioinformatics and Genomic Medicine" aims to address all the bioinformatics aspects necessary for the production and interpretation of Next Generation Sequencing (NGS) data within a clinical genetics laboratory with examples of rare diseases and oncogenetics (<https://www.fun-mooc.fr/en/courses/big-introduction-bioinformatics-genomic-medicine/>) (Figure 8).

This introductory course was intended for health professionals using genomics. Its objective is to provide specific and adapted content to enable them to understand the different steps from phenotyping to molecular diagnosis and to have a critical eye on the analyses while considering the pitfalls and limits of NGS.

Each teaching unit explored a step of NGS processing by focusing on different themes with videos, texts, and self-correction exercises. Interactive content, such as



Python-based Jupyter notebooks, permitted to go further in genome interpretation and programming.

Sixteen teachers from 11 institutes participated in the making of educational units. European volunteers from the French Medical Genetics Resident Society (SIGF, <https://interne-genetique.org/>) and ESHG Young (<https://www.eshg.org/index.php?id=eshgy>) were crucial in proofreading to ensure the MOOC clarity. Nearly 12,000 learners in two years subscribed to the course from 134 different countries, providing a global learning resource in genomic medicine. This MOOC was the subject of one master thesis and one MD thesis.

The MOOC B.I.G. initiative was financed by the AnDDI-rare healthcare pathway (Health Sector Developmental Disabilities with or without Intellectual Disability of Rare Causes, <http://anddi-rares.org/>), the ERN ITHACA (European Reference Network for Rare Malformation Syndromes, Intellectual and Other Neurodevelopmental Disorders, <https://ern-ithaca.eu/>), and SFMPP (French Society of Predictive and Personalized Medicine, <https://www.sfmpp.org/>). The realization and hosting were supported by the MOOC factory of the Center for Interdisciplinary Research (CRI, <https://cri-paris.org/>) and the Université Numérique en Santé et Sport (UNESS, <https://www.uness.fr/>).

- › Bienvenue à bord !
- › Welcome aboard !
- › Forum
- › Retour aux sources
- › Unité 1 - Comprendre la révolution du SHD en génétique médicale
- › Unité 2 - Interprétation d'une variation génétique
- › Unité 3 - Appel des variations génétiques
- › Unité 4 - Alignement sur le génome de référence
- › Unité 5 - Appel des bases du séquençage
- › Avant de partir...
- › Back to basics
- › Unit 1 - Understanding the NGS revolution in genomic medicine

VOIR L'UNITÉ DANS STUDIO

←
📄
📄
📄
☰
→

## ANNOTATION TOOLS

Given the amount of information needed to interpret a variant, these aggregation interfaces are a valuable aid to access all these annotations in the blink of an eye.

← TIP: Multiple pages to read! →

### KEY POINTS

- The annotation of the variants provides the information necessary for their interpretation.
- Graphical tools for variant by variant analysis and command line tools for large-scale annotation are available.
- Updating it with the latest scientific advances makes it possible to find new diagnostics.
- A regular update of the annotations is recommended for a precise diagnosis.

**KÉVIN YAUY**

Medical Geneticist  
Montpellier University  
Hospital (CHU)

🗣️ Activate English subtitles with the 'cc' button on the video navigation bar 🕒

Figure 8. A screenshot of the MOOC BIG.

## 3-minutes Thesis

An essential part of a research/thesis project is communication about the topic. It helps to have feedback on one's work to improve the storytelling and get a better impact on the community. Part of this communication process was my participation in the French 3-minutes thesis competition in 2021. After a workshop on Scientific communication provided by Université Grenoble Alpes and Ludovic LECORDIER from "Spontanez-vous" (<https://spontanez-vous.fr/>), I was selected as a Finalist in the French Alps (Figure 9). Here's the presentation I performed on March 9th, 2021.

[Speech transcript translated from French]

*"Genetics on a sling*

*Chromosomes in the atmosphere*

*Taxis to the galaxies*

*And my flying carpet?*

[Translated lyrics from French by Noir Désir song, *Le vent nous portera*]

*Don't you find these words a bit suspicious between you and me? So I'm not on any substance, but dear listeners of 3-minutes Thesis FM Radio, I take control of the radio station so we can try to see things more clearly together.*

*My name is Kevin, I'm a medical geneticist, and I'd like to tell you about my breadcrumb trail. I am dedicated to exploring the human genome in search of a lucky star, a diagnosis for my patients with rare diseases.*

*So get ready. We're going to do a little flashback. Here we are in 2001, "Le vent nous portera" had just been released, and at the time, we hadn't even finished sequencing the first human genome...*

*And twenty years later, we have the incredible ability to read all the letters that make up our genetic heritage in just a few days. The genome is like a library of recipes that allow you and me to build ourselves as human beings. But unfortunately, sometimes, spelling mistakes can cause rare diseases. These spelling mistakes are called variants in our jargon. And to find the variant that causes my patient's disease, in the 3 billion letters of the genome, heeeee.... is a bit complicated.*

*I had to learn to rely on a machine; I had to learn to code, write computer programs, and be a geek! Well, you'll tell me I already have the look. All I had to do was to be able to code a compass that would guide me through the genome to solve too-long diagnostic odysseys.*

*After learning a second job, I realized that it is sometimes easier to talk to a machine than to a human being, so I created an artificial intelligence that I summarized in four words. "Back to the future" because it uses today's knowledge to solve yesterday's enigmatic cases.*

*Genome Alert! identifies relevant new information from the literature to guide a journey to old, unsolved genetic analyses. And attempts to change the present by automatically targeting new potential diagnostic variants.*

*And hold onto your hats. With a wave of the magic wand, Genome Alert! has scanned over 5000 analyses, solved the diagnostic puzzle of at least six patients, and changed their management. And that's just the beginning.*

*So dear listeners, the wind carried me towards a desire, I would say, unexpected to IT. For this thesis, I put away my trusty stethoscope and learned to rely on my computer. And I believe I will continue to do so that none of my patients will remain without an answer/diagnosis."*

[End of speech transcript]

The intervention (in French) is available at this link:

<https://youtu.be/7bDEPShzxp4?t=4475>.



Figure 9. French 3-Minutes Thesis competition poster of the French Alps Final.

## Chapter III - the reinterpretation challenge

### *Diving into Genome Alert!*

As sequencing is no longer limited in genomic medicine, one of the main challenges in rare disease diagnosis is the interpretation and iterative re-interpretation of the multitude of variants detected. Indeed the diagnostic yield of NGS depends on clinical entities, but globally, a majority of patients were undiagnosed after sequencing <sup>26</sup>. Studies have already reported that reanalysis of previous genomic analysis could significantly improve diagnostic yield. However, the reinterpretation task was said to be highly manual, time-consuming, and primarily uncovered in healthcare systems <sup>3,52</sup>. There is a need for guidelines in variant reinterpretation that can facilitate implementing a low-cost, scalable, and accessible approach in genomic centers worldwide <sup>53</sup>. If progress has been made to automate genomic variant interpretation, the American Society of Human Genetics statements reinforce the need for a standardized approach to genomic reanalysis <sup>54</sup>.

I first described how the sequence variant interpretation is performed to provide more context on this challenge we were trying to tackle.

## The variant interpretation challenge

NGS detects a significant amount of genetic variation, approximately 20,000 SNV per individual in the coding genome. If a Mendelian genetic disease is suspected (i.e., a variant or a few variants cause the patient's disease), the interpretation of constitutional variants corresponds to identifying the variant(s) of interest amongst this considerable mass of data.

### *Discrepancies and Subjectivity of Interpretation*

Since the early 2010s, NGS has been used for medical diagnostics, especially in the field of rare diseases and oncology. However, significant differences in variant interpretation have been reported between different testing centers and genetic centers <sup>47</sup>. The complexity of the interpretation process can account for these differences: the combination of multiple sources of evidence relating to clinical data, biological data, population genetic data, etc... To manage this, the genetics community has established and approved guidelines, specifically the ACMG-AMP 2015 guidelines <sup>48</sup>. These describe different criteria used as evidence when interpreting genetic variants and the weighting for each piece of evidence.

### *A Bundle of Arguments to Question*

Features used in variant interpretation include several categories: clinical relevance of the gene, the molecular impact of a variant, segregation of the variant to affected and unaffected relatives in a family, and functional studies. First, the clinical relevance of a genetic variant is assessed by comparing the patient's phenotype (or



symptoms) to features known to be associated with the disease caused by the gene. This is done by consulting the medical literature. As a result, detailed phenotypic data allows for more accurate variant interpretation.

Next, the molecular impact of a variant on the protein and gene function is determined by several arguments: in silico computational predictions about the variant's effect on protein structure, evolutive conservation across species of the amino acid, population frequency of the variant, or reports of this variant in other affected individuals.

Finally, in some cases, the search for the variant in other family members (segregation study) or an extensive functional analysis (in vitro studies) of the variant must be conducted to make a decision.

#### *The Outcome of Interpretation: a 5 Tier Classification System*

After combining all the evidence, the final aim of variant interpretation is to classify the variant into one of five classifications. These classifications provide a standard communication method between clinicians and scientists and were rapidly applied by a large part of the medical community worldwide. They indicate the criteria to be evaluated in the interpretation process, the weight to be given to them, and the algorithm for assigning them a classification from class 1 (benign variant) to class 3 (a variant of uncertain significance) to class 5 (pathogenic variant). The criteria can be in favor of pathogenicity (P) or favor of benignness (B). Criteria are weighted according to the level of confidence: stand-alone or absolute (A), very strong (VS), strong (S), moderate (M), or supporting (P). It should be noted that despite the impossibility of quantifying the uncertainty related to the interpretation of the

variants, classes 2 (likely benign variant) and 4 (likely pathogenic variant) correspond to a probability of more than 90% that the variant is benign or pathogenic, respectively.

To summarize, sequencing and detecting variants is no longer the limiting factor in the NGS era. The real challenge lies in interpretation. A combination of evidence is required to conclude that variants are pathogenic. To standardize the interpretation of variants, recommendations exist, such as those proposed by the ACMG-AMP. Unfortunately, in most cases, the variant impact remains of uncertain significance.

## Data sharing, a key element in genomic medicine

As variant interpretation is a challenging task, sharing data becomes a central element in genomic medicine, as it allows us to benefit from the dynamics of scientific publication (slow) and the diagnostic progress of the international community (fast).

### *Salvation from data sharing*

To facilitate variant interpretation, it is necessary that the available data is shared and the interpretations made are also shared to solve diagnoses and discover new genotype-phenotype correlations <sup>49</sup>. Using an interpretation already performed by other biologists saves time for the patient. Sharing data is even more helpful when a

variant is rare or difficult to interpret. This is especially true for variants involved in very rare diseases where the amount of data is small <sup>50</sup>. There are several initiatives in this area, and the current most active resource in genomic medicine is ClinVar.

### *ClinVar, a community-driven database*

ClinVar, supported by the National Institutes of Health (NIH), is a free, public database listing interpretation of known variations with clinical information and criteria useful for interpretation <sup>51</sup>. The variants are classified according to the ACMG recommendations and the level of evidence provided. It will have a star rating: ratings range from 0 stars (little or no documented methodology) to 4 stars (practice guidelines). Over one million variants interpreted in a clinical context are available. It is one of the gold standard resources for medical genetics.

## Motivation

As a physician, I was frustrated by not being able to provide reinterpretations for unsolved patients. I put a lot of effort into asking clinical laboratory specialists to perform reinterpretation only when patients needed it urgently. With the current technology available, I couldn't understand why we didn't yet have access in hospitals to a semi-automated system to help us monitor new clinical knowledge in variant classification and alert us when it could change the patient's diagnosis.

We bet exploiting data-sharing databases could provide a method to standardize genomic analysis reinterparation and supply a scalable and affordable system that the community can adopt. To prove it, we decided to take advantage of the database ClinVar. ClinVar's highly accessible collaborative platform (<https://www.ncbi.nlm.nih.gov/clinvar/>) is widely recognized as one of the most dynamic genomic databases<sup>55</sup>. ClinVar is updated weekly with thousands of changes and additions that can impact diagnostic performance. Surprisingly, ClinVar doesn't provide the history of classification changes in the database, and no tools were available to do it either.

Overall, we described in a scientific article the development and evaluation of a semi-automated method for reassessing variant pathogenicity and genotype-phenotype knowledge in the ClinVar database called *Genome Alert!* that solves numerous diagnostics. This study is published in the Genetics in Medicine journal with an open access <sup>56</sup> ([https://www.gimjournal.org/article/S1098-3600\(22\)00654-2/fulltext](https://www.gimjournal.org/article/S1098-3600(22)00654-2/fulltext)). A webapp to use this method is accessible at <https://genomealert.univ-grenoble-alpes.fr/> and open source code at <https://github.com/SeqOne>.

I was the principal investigator of the scientific project and coordinated this work with our collaborators (Eurofins Biomnis, Cerba, and CHU de Rouen). With the help of Jérôme Audoux, Sacha Beaumeunier, Nicolas Soirat, Abdoullaye Diallo, Raphael Lanos and Melanie Broutin from SeqOne Genomics, I programmed scientific experiments, scripts and webapp to make these methods accessible to the


community. I supervised Quentin Fort, who participated in scientific experiments during his internship at SeqOne Genomics. Julien Thevenon supervised me for the manuscript writing.



## ARTICLE

# Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene–phenotype reassessment in clinical routine



Kevin Yaou<sup>1,2,\*</sup> , François Lecoquierre<sup>3</sup>, Stéphanie Baert-Desurmont<sup>3</sup>, Detlef Trost<sup>4</sup>, Aicha Boughalem<sup>4</sup>, Armelle Luscan<sup>4</sup>, Jean-Marc Costa<sup>4</sup>, Vanna Geromel<sup>5</sup>, Laure Raymond<sup>5</sup>, Pascale Richard<sup>6</sup>, Sophie Coutant<sup>3</sup>, Mélanie Broutin<sup>2</sup>, Raphael Lanos<sup>2</sup>, Quentin Fort<sup>2</sup>, Stenzel Cackowski<sup>7</sup>, Quentin Testard<sup>1,5</sup>, Abdoulaye Diallo<sup>2</sup>, Nicolas Soirat<sup>2</sup>, Jean-Marc Holder<sup>2</sup>, Nicolas Duforet-Frebourg<sup>2</sup>, Anne-Laure Bouge<sup>2</sup>, Sacha Beaumeunier<sup>2</sup>, Denis Bertrand<sup>2</sup>, Jerome Audoux<sup>2</sup>, David Genevieve<sup>8</sup>, Laurent Mesnard<sup>9,10</sup>, Gael Nicolas<sup>3</sup>, Julien Thevenon<sup>1</sup>, Nicolas Philippe<sup>2</sup>

### ARTICLE INFO

#### Article history:

Received 30 November 2021

Received in revised form

7 February 2022

Accepted 7 February 2022

Available online 17 March 2022

#### Keywords:

ClinVar

Gene–phenotype associations

Sequencing reinterpretation

Variant pathogenicity

### ABSTRACT

**Purpose:** Retrospective interpretation of sequenced data in light of the current literature is a major concern of the field. Such reinterpretation is manual and both human resources and variable operating procedures are the main bottlenecks.

**Methods:** Genome Alert! method automatically reports changes with potential clinical significance in variant classification between releases of the ClinVar database. Using ClinVar submissions across time, this method assigns validity category to gene–disease associations.

**Results:** Between July 2017 and December 2019, the retrospective analysis of ClinVar submissions revealed a monthly median of 1247 changes in variant classification with potential clinical significance and 23 new gene–disease associations. Re-examination of 4929 targeted sequencing files highlighted 45 changes in variant classification, and of these classifications, 89% were expert validated, leading to 4 additional diagnoses. Genome Alert! gene–disease association catalog provided 75 high-confidence associations not available in the OMIM morbid list; of which, 20% became available in OMIM morbid list. For more than 356 negative exome sequencing data that were reannotated for variants in these 75 genes, this elective approach led to a new diagnosis.

**Conclusion:** Genome Alert! (<https://genomealert.univ-grenoble-alpes.fr/>) enables systematic and reproducible reinterpretation of acquired sequencing data in a clinical routine with limited human resource effect.

© 2022 The Authors. Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Gael Nicolas, Julien Thevenon, and Nicolas Philippe jointly supervised this work.

\*Correspondence and requests for materials should be addressed to Kevin Yaou, Institute for Advanced Biosciences, UGA/Inserm U 1209/CNRS UMR 5309 joint research center, Site Santé-Allée des Alpes, 38700 La Tronche, France. *E-mail address:* [kevin.yaou@univ-grenoble-alpes.fr](mailto:kevin.yaou@univ-grenoble-alpes.fr)

Affiliations are at the end of the document.

doi: <https://doi.org/10.1016/j.gim.2022.02.008>

1098-3600/© 2022 The Authors. Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## Introduction

Genetic tests are increasingly prescribed and included in health care pathways for diverse clinical indications.<sup>1,2</sup> Several countries have developed population genomics organizations that are revolutionizing medical practices.<sup>3,4</sup> However, many of these genomic analyses remain inconclusive owing to limitations in genomic and medical knowledge available at the time of analysis.

The American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) recommendations for variant classification aim at standardizing variant interpretation practices in genomic centers, in the context of medical interpretation.<sup>5</sup> Recently, tools have been published to automatically classify genomic variants on the basis of these recommendations.<sup>6-8</sup> Meanwhile, evolving medical knowledge and rapid adoption of clinical genome sequencing have influenced the standard practices and have created additional needs. A current and major preoccupation in this field is the definition of standards for periodic and prospective reanalysis of existing sequencing data. Indeed, reanalyzing existing genomic data improves diagnostic yield (7% increase per year).<sup>9,10</sup>

In practice, such an in-depth reinterpretation is mainly manual and time-consuming, with major bottlenecks such as human and funding resources or lack of consistency between centers. Clinical recommendations from the American and European Societies of Human Genetics reinforce the need for a standardized and automated approach to the reinterpretation of genomic analyses.<sup>11-14</sup> Some companies offer paid black box services, with poorly detailed methods that cannot be reproduced.<sup>15,16</sup>

Clinical knowledge of rare diseases is contained in expert-curated databases (such as OMIM<sup>17</sup> or Clinical Genome Resource [ClinGen]<sup>18</sup>), peer-reviewed medical literature, and information sharing between health practitioners through community-based platforms (such as MatchMaker Exchange<sup>19</sup> or ClinVar<sup>20</sup>). Reliability and exhaustiveness of information vary widely across these data sources. Furthermore, careful monitoring of clinical knowledge by every laboratory represents an organizational challenge for a prospective reanalysis of acquired data. To enable a systematic, reproducible, and prospective genome interpretation, a collaborative approach for clinical knowledge aggregation combined with automated medical knowledge monitoring and curation is needed.

The main community-based repository of genomic knowledge is ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), a shared variant interpretation database that featured 1 million submissions in 2020. ClinVar is updated weekly with several thousands of modifications of variant classifications that could affect the diagnostic yield of previous analyses. There is currently no monitoring system that can highlight these changes at a scale for the complete database. Besides variant classification, gene-phenotype

association catalogs are crucial because they are commonly used to design phenotype-specific gene panels for dry-lab filtering and set the frontiers for clinical genome analysis.<sup>21,22</sup> Although not their primary purpose, variant-centered databases could also theoretically provide a complementary resource to gather gene-phenotype knowledge.

In this article, we detail an automated method for the reassessment of variant pathogenicity and gene-phenotype associations through ClinVar follow-up. This procedure, called Genome Alert!, aims at performing a routine and systematic reinterpretation of existing genomic data. The procedure's effectiveness was evaluated through a 29-month multicentric series (2018-2019) of 5959 consecutive individuals screened using targeted sequencing (4929 individuals with hereditary cancers) and exome sequencing (1000 analyses including 356 undiagnosed individuals with suspected Mendelian disorders).

## Materials and Methods

### Genome Alert! standardized procedure

ClinVCF, Variant Alert!, and ClinVarome are a suite of tools that constitute the heart of the Genome Alert! standardized procedure.

### ClinVCF: A ClinVar quality processing method

Before comparing different versions of the same source, data consistency needs to be verified. This first step is based on ClinVCF tool, and once every submission has been tracked, data will be processed for the next step.

ClinVCF imports monthly updated ClinVar Xtensible Markup Language (XML) files. XML format was preferred over VCF mainly because of better consistency and traceability across versions for the ClinVar Variation ID, the history of changes in each variant classification, and the additional gene-phenotype data availability in XML. ClinVCF considers an automatic reclassification of variants with at least 4 submissions and conflicting interpretations of pathogenicity status. Consensus classification according to ClinVar policies sets the conflicting interpretations of pathogenicity status when at least 1 conflict in submission is observed, except if an expert consortium (as ClinGen) has defined classification (details available in [Supplemental Method 1](#)). On the basis of the provided classifications transformed from literal transcription (eg, likely pathogenic) to class number (eg, class 4), if  $\geq 4$  submissions are available, a new consensus is proposed after outlier submissions removal according to the 1.5\* Interquartile Range (IQR) Tukey method.<sup>23</sup> We only reclassify variants from conflicting status to likely pathogenic or pathogenic status. ClinVCF provides a 3-tier reclassification confidence score detailed in [Supplemental Figure 1](#). As an output, ClinVCF writes a Variant Calling File (VCF) v4.2 file.

### Variant Alert!: A variant knowledge monitoring tool

Variant Alert! tool aims at identifying changes in variant classification across 2 versions of the database. Changes were defined as (1) a modification in the classification of an existing variant and (2) the creation or suppression of a variant entry.

Stratification of the consequences in classification modification was proposed (Supplemental Table 1). Major classification modification was defined as a change that may affect the clinical management of a patient (eg, uncertain significance to likely pathogenic status). Minor classification modification was defined as a change that may not affect the clinical management of a patient (eg, pathogenic to likely pathogenic status).

Variant Alert! writes 2 files: (1) the list of variants that were modified, added, or removed and (2) the list of genes that were added to or removed from the database. This gene list is notably used by ClinVarome.

### ClinVarome: A method for automated gene–disease association evaluation

ClinVarome tool aims to periodically and automatically evaluate gene–disease association in the ClinVar database. To differentiate genes on the basis of their clinical validity, the work from European Molecular Biology Laboratory–European Bioinformatics Institute Gene2Phenotype,<sup>24</sup> ClinGen,<sup>18</sup> and Genomic England PanelApp<sup>25</sup> were first compared. Although theoretically comparable, their rationales and contents were partially overlapping and with conflicting classifications. To discriminate candidate genes from definitive gene–disease associations, we decided to use an unsupervised clustering model. Only the genes with at least 1 likely pathogenic or pathogenic variant (single nucleotide variant or indel affecting a single gene) in ClinVar were considered in a list called ClinVarome. As a consensus criterion, we chose to assess the strength of a gene–disease association through the quantification of 4 variables: (1) count of likely pathogenic and pathogenic variants, (2) highest variant classification (CLNSIG, likely pathogenic or pathogenic), (3) highest ClinVar review variant confidence (CLNREVSTAT, from 0 to 4 stars), and (4) time interval between the first and the last pathogenic variant submission (replication of the gene–disease association event). For these 4 variables, values were gathered through periodic monitoring of changes in the database following the ClinVCF and Variant Alert! tool procedures. Clustering variants according to these variables allowed us to define clusters of genes according to their clinical validity. The scikit-learn Agglomerative Clustering tool (parameters: Euclidean affinity, ward linkage) was used, and t-distributed stochastic neighbor embedding representation (parameters: 2 components, perplexity 150, 2000 iterations, and 1000 iterations without progress) was performed. Gene–disease validity classification was computed per gene but not per disease. The Gene Curation Coalition (GenCC) (<https://thegencc.org/>) database was released recently and was used to evaluate ClinVarome. To compare ClinVarome

clusters and GenCC classification, GenCC submissions were summarized into 3 categories (Green, Orange, Red) (Supplemental Methods 2).

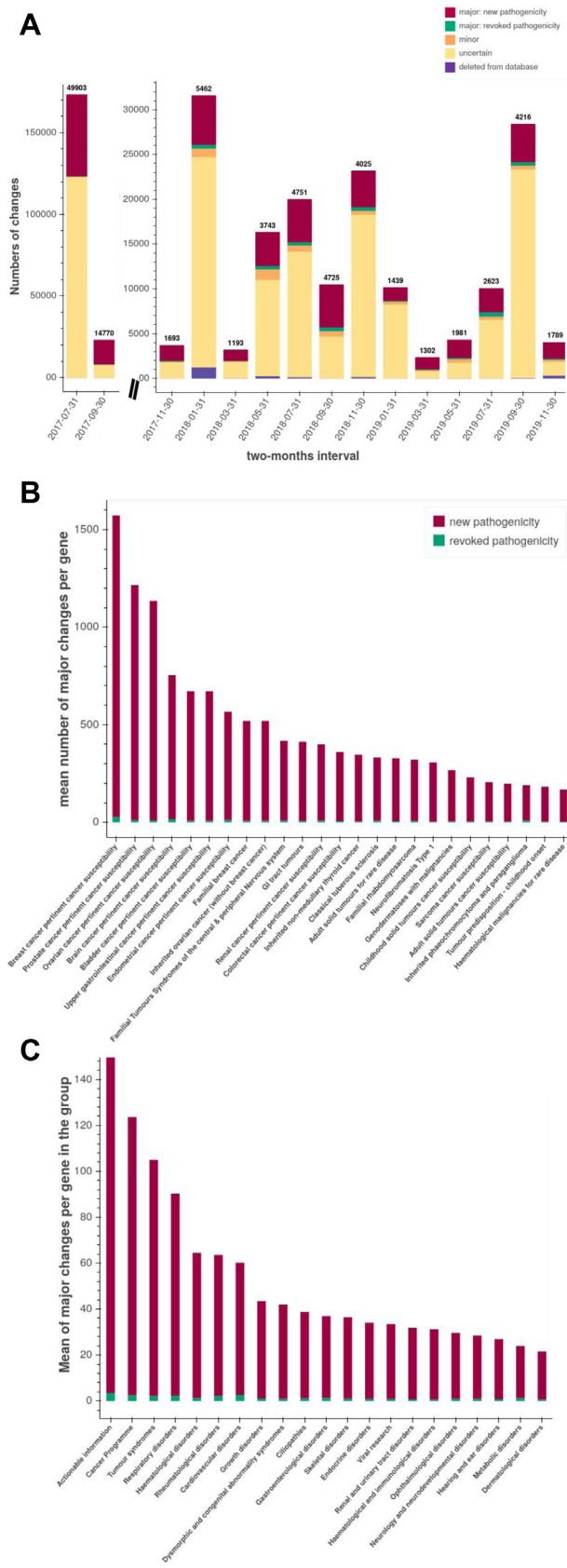
### Study design and participants

To evaluate the clinical impact of Genome Alert!, we collected 5929 consecutive germline sequencing data samples from 3 centers in France between July 2017 and December 2019 as part of their routine genetic investigation: (1) a variant database gathering all class 3 (uncertain significance), class 4 (likely pathogenic), and class 5 (pathogenic) variants identified in a colon cancer–targeted sequencing (14 genes) sequenced in 2540 individuals in the Rouen University Hospital; (2) a cancer-targeted sequencing data set of 2389 individuals by the Cerba laboratory (66 genes); and (3) exome sequencing data of individuals with developmental disorders, rare kidney diseases, or other rare diseases as follows: 108 probands from the Rouen University Hospital, 477 probands (with 356 negative analysis) from the Cerba laboratory, and 415 probands from the Eurofins Biomnis laboratory. Patient samples, together with a basic phenotype description and molecular diagnosis (when available), were anonymized. Two main clinical evaluations were performed: (1) variant-centered reanalysis, which aims at matching individuals that carry exact variants with potential clinical significance reported by Genome Alert!, and (2) gene-centered reanalysis, which aims at matching individuals who carry candidate variants in high-confidence clinical genes referenced in ClinVarome and not in OMIM. Initial analyses were performed between 0 and 2 years before this reanalysis.

### Selection of variants with potential clinical significance

All sequencing data were systematically reinterpreted according to Genome Alert!'s report and compared with the initial variant interpretation. For targeted sequencing and exome reanalysis, genomic positions of variants with major changes in classification were queried in the existing patient's variant calling files (variant-centered analysis). For exome data, we performed a reanalysis of variants in VCF with the following criteria: (1) among 75 ClinVarome morbid genes, which were not available in OMIM, and with a second event of gene–disease validation (including a likely pathogenic or pathogenic variant with ClinVar review confidence  $\geq 2$  stars and a likely pathogenic or pathogenic variant entry subsequent to the initial entry); (2) variant not shared with another individual in the series; (3) sufficient sequencing quality (variant allele fraction  $> 25\%$  and read depth  $> 20$  reads); (4) rare in Genome Aggregation Database<sup>26</sup> population (frequency  $< 10^{-5}$  if heterozygous genotype or  $10^{-4}$  if homozygous genotype); and (5) protein consequence among nonsense, frameshift, missense (missense are selected with Combined Annotation





**Figure 1 ClinVar variant classification monitoring between July 2017 and December 2019.** A. Bar chart distribution of every 2 months of changes in variant classification. The bar chart was

Dependent Depletion<sup>27</sup> score > 30 and MetaSVM<sup>28</sup> = D), or splice variants (based on dbcsnv RF<sup>29</sup> predicted impact score > 0.6) (gene-centered reanalysis).

## Results

### ClinVar knowledge dynamics

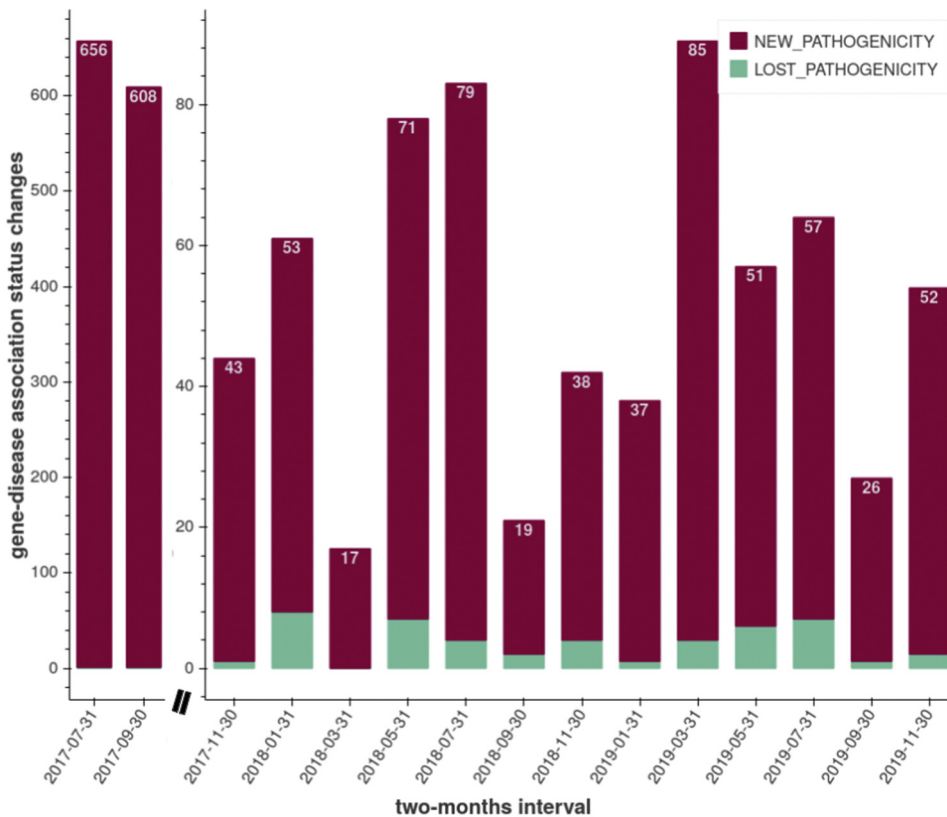
To get insights into variant classification and gene–disease association and to estimate the amount of new clinically relevant information in the ClinVar database available through time, a retrospective analysis of ClinVar submissions over 29 months was performed (July 2017 [included] to December 2019). Of note, VCF genomic positions in ClinVar were introduced in July 2017 and probably are associated with the largest injection in the ClinVar database.

The number of variants with ACMG/AMP classification<sup>5</sup> increased from 144,943 to 491,838. Among modifications in the database, the count of major changes was 107,167 in ACMG/AMP classification, and among these, 103,615 resulted in a pathogenicity status, which was previously unreported, whereas 3552 resulted in the revocation of a previously established pathogenicity (Figure 1A). These changes varied significantly according to disease groups the between gene panels (according to Genomics England PanelApp), in which the oncogenetic panels were on top of the list of panels. The panels and disease groups presenting most of the changes per gene are presented in Figure 1B and C and Supplemental Table 2. Clinical gene entries in ClinVar were also monitored. A median of 23 ClinVar morbid genes per month that were newly associated with Mendelian disease was observed (Figure 2).

### Changes in variant classification

To evaluate the robustness of clinical variant information, the consistency of variant classification was explored and is described in Supplemental Table 3. Among 144,943

split for better readability. Bold numbers and dark red color represent new (likely) pathogenic variant entries, green represents number of revoked (likely) pathogenic variants, orange represents number of minor change variants (eg, pathogenic to likely pathogenic), yellow represents number of changes with uncertain clinical impact (VUS or conflict entry), and purple represents number of changes leading to variant disappearance. B. Bar chart of top panels with clinically significant changes per gene (major changes). Dark red color represents (likely) pathogenic variant entries, and green represents revoked (likely) pathogenic variants. C. Bar chart of top disease group with clinically significant changes per gene (major changes). Dark red color represents (likely) pathogenic variant entries, and green represents revoked (likely) pathogenic variants. GI, gastrointestinal tract; VUS, variant of uncertain significance.



**Figure 2** ClinVar clinical genes entries associated with new or deprecated Mendelian disease (morbid status) distribution between December 2017 and December 2019. The bar chart was split for better readability. Dark red represents morbid genes entries (first variant with likely pathogenic or pathogenic status), and green represents revoked morbid genes. White numbers represents number of new morbid gene entries by 2 months.

variants available in July 2017, 10,254 (7%) were reclassified between July 2017 and December 2019, ie, we observed only a small portion of variants being reclassified over time. These reclassifications included automatically reclassified variants with conflicting interpretations. More precisely, among the 11,417 likely pathogenic variants, 1125 (9.94 %) variants were reclassified as benign variants, likely benign variants, variants of uncertain significance, or variants with conflicting interpretations of pathogenicity.

### Automatic variant reclassification with conflicting interpretations

A criticism of the ClinVar database is the misclassification of pathogenic variants, such as the well-known *HFE* pathogenic variant NM\_000410.3:c.845G>A. We observed that it was mostly due to a unique outlier submission with a classification for a distinct condition (eg, cutaneous photosensitivity porphyria phenotype). We evaluated our method to remove such outlier submissions. Among all the variants available in ClinVar in December 2019, 22,973 of a total of 503,994 (4.5%) variants were classified with a conflicting interpretation of pathogenicity. Genome Alert! automatic reclassification method proposes to detect outlier submissions to suggest a consensus classification. This

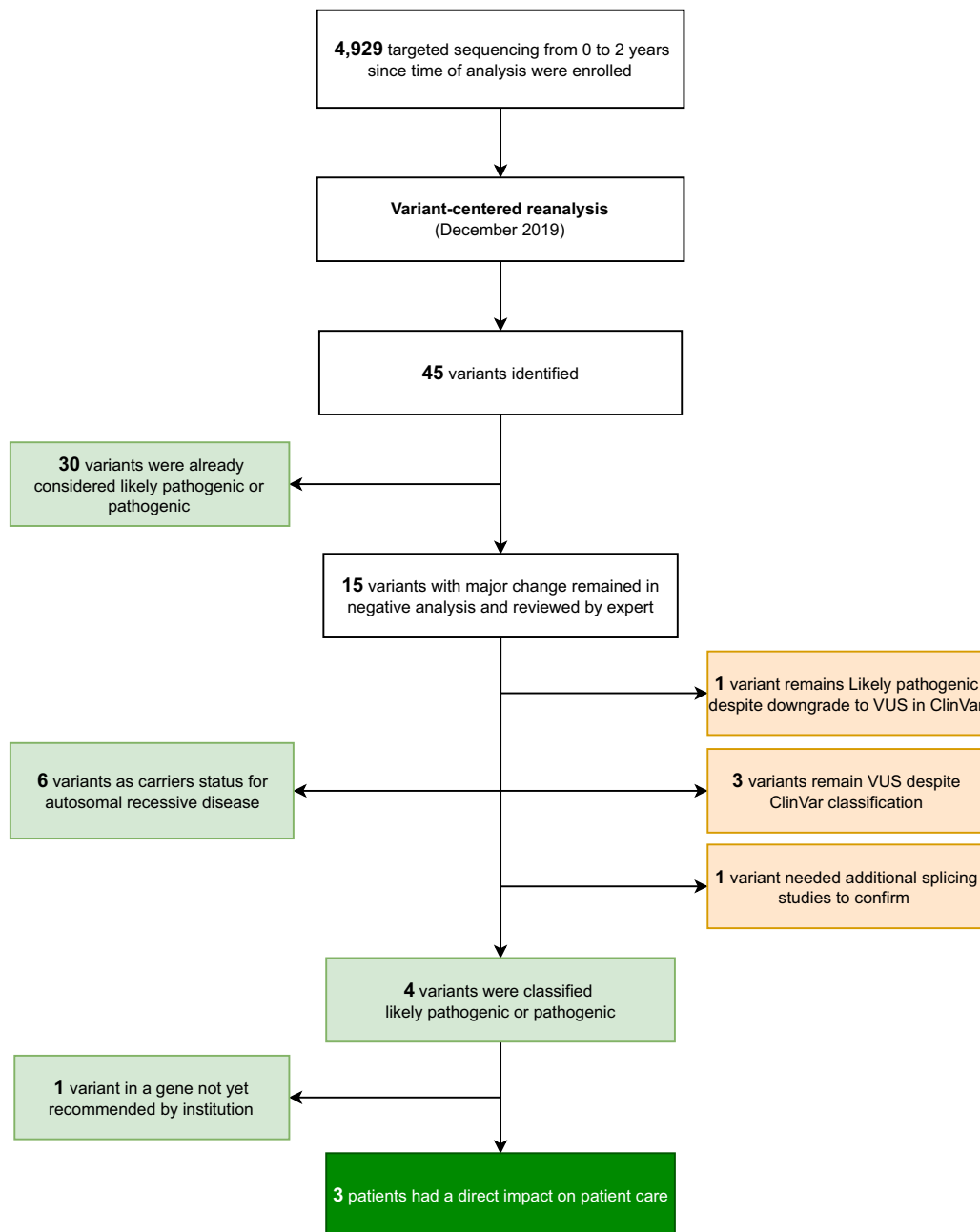
allowed the reclassification of 188 variants from conflict to likely pathogenic or pathogenic classification in 135 genes and 1625 variants in 436 genes from conflict to likely benign or benign classification (Supplemental Table 4, Supplemental Figures 1 and 2).

Variants automatically reclassified as likely pathogenic or pathogenic in cancer ( $n = 9$ ) and cardiogenetic disease ( $n = 11$ ) were presented to French National experts in the field. Of these 20 automatic reclassifications, 17 were confirmed as accurate by experts and 3 remained as variants of uncertain significance, lacking evidence of pathogenicity for our experts.

### Clinical impact of changes in variant classification

To assess the clinical impact of Genome Alert!'s changes in variant classification, previously analyzed cancer-predisposition targeted sequencing data were assessed (4929 individuals from 2 genetic centers) (variant-centered reanalysis, Figure 3). Among all variants detected in this cohort, this method highlighted 45 variants with major changes between the time of analysis and December 2019, which were proposed for manual review by their referring geneticists (Supplemental Tables 5 and 6).

Among the 45 variants, 30 had been already manually reported by the clinical geneticists as likely pathogenic or



**Figure 3 Experimental design of the variant-centered reanalysis.** Flow charts describing how the sequencing data were reinterpreted according to variant reclassification only. Green box represents new diagnosis. Light green boxes represent confirmed variant classification. Orange boxes represent excluded variants. VUS, variant of uncertain significance.

pathogenic at the initial time of analysis, meaning that these classifications were ahead of the ClinVar database. The 15 unreported variants were manually curated, looking for additional diagnoses. Among them, 14 variants were newly classified as likely pathogenic or pathogenic and 1 was downgraded as a variant of uncertain significance (VUS) in ClinVar. The manual curation of these 14 variants led to the conclusion that 6 corresponded to a carrier status for a recessive disorder, 3 were manually classified as VUS, and 5 were submitted to a multidisciplinary meeting for external review. Finally, 4 of these latter 5 were classified as likely

pathogenic or pathogenic by experts leading to additional diagnoses. One variant remained classified as a VUS, and complementary studies on the patient's messenger RNA were proposed before conclusion (*PALB2*, NC\_000016.9(NM\_024675.3):c.3350+4A>G). Finally, an 89% validation rate (40 of 45) of major changes were observed. This variant reclassification tracking system allowed an additional diagnosis per 1000 analyses.

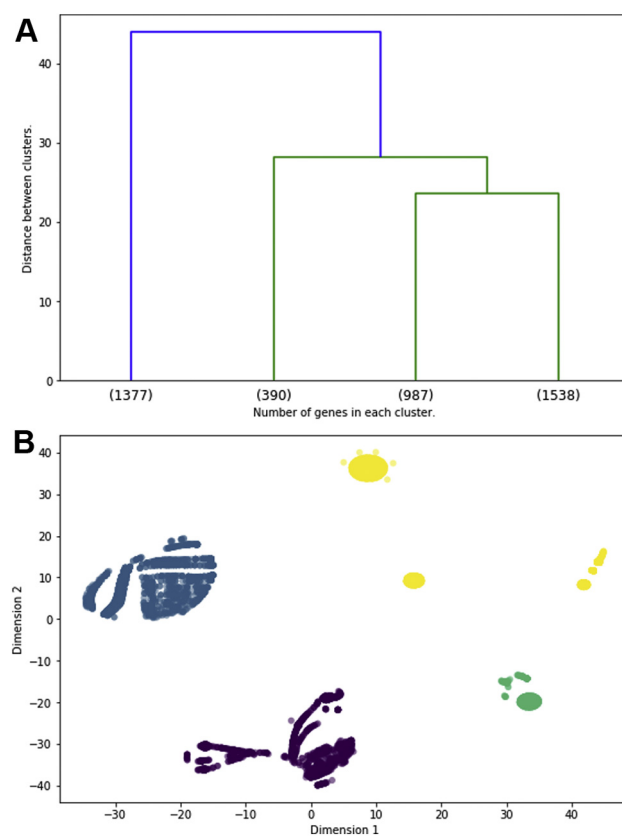
Replication of the variant-centered reanalysis was performed in the exome sequencing cohort, looking for variant exact match. Selective reanalysis in previous exome

sequencing analysis (1000 individuals in 3 genomic centers) highlighted <1 variant per exome (only 297 variants) with major changes between the time of analysis and December 2019. These 297 variants were then explored by clinical geneticists. Among all 297 variants, 1 variant (*POLG*, NM\_002693.2:c.2243G>C) was automatically reclassified as pathogenic by our IQR outlier submission method and was initially reported as VUS, thus helping us to confirm the diagnosis. Compound heterozygosity was observed for a pathogenic variant (*POLG*, NM\_002693.3:c.1399G>A). Exome sequencing reanalysis with the variant-centered reanalysis also provides an additional diagnosis per 1000 analyses.

### Monitoring ClinVar gene–disease association knowledge

A focus has been toward exploring rarely explored gene–disease association in ClinVar data. To discriminate candidate genes from definitive gene–disease associations in ClinVarome, unsupervised clustering was performed on the basis of the following criteria: (1) count of likely pathogenic and pathogenic variants, (2) highest variant classification, (3) highest ClinVar review variant confidence, and (4) time interval between the first and the last pathogenic variant submission. According to distances between clusters and model dendrogram, the number of clusters was set to 4 (Figure 4). Careful observation of these clusters identified objective patterns to understand the classification. We observed that all genes in the first and second clusters had a reproducibility event (a new likely pathogenic or pathogenic variant entry, the confirmation of the likely pathogenic or pathogenic classification by another submitter or expert panel) in pathogenicity status, thus giving them strong confidence. Genes from the first cluster hold pathogenic variants with ClinVar's  $\geq 2$  stars of review confidence and the second cluster genes include pathogenic variants with different entry dates and <2 stars of review confidence. Genes in the third cluster had 1 strong argument for pathogenicity but needed another event to be fully confirmed (the third cluster genes contained at least 1 pathogenic variant and all pathogenic entries were added at the same date). Because genes in the fourth cluster were only likely pathogenic variants, their gene–disease association remained to be confirmed (Supplemental Table 7).

To assess the exhaustivity of the ClinVarome, a comparison with the OMIM database was performed. In December 2019, there was a 95% overlap (3675/3858) between OMIM morbid clinical genes and ClinVarome morbid genes. Overall, 365 genes were referenced only in OMIM and not in ClinVarome. We observed patterns that were not available in ClinVar. These patterns include nonconfirmation of a disorder as a genuine Mendelian disorder (only 1 publication or isolated patient reports), susceptibility to multifactorial disorders or infection, referencing of genes belonging to



**Figure 4 ClinVarome morbid genes exploration and gene–disease validity classification.** A. Agglomerative clustering dendrogram of ClinVarome in December 2019. B. t-distributed stochastic neighbor embedding representation of ClinVarome 4 variables by gene data. Green represents fourth cluster (390 genes), yellow represents third cluster (987 genes), blue represents second cluster (1538 genes), and purple represents first cluster (1377 genes).

molecular mechanism distinctive from a single gene disorder as microdeletion or microduplication syndromes, Mendelian traits that are not diseases, epigenetic loci, genes with targeted pathogenic complex variants, and very recently described diseases. The evaluation focused on these 519 specific genes, referenced only in ClinVar and not in OMIM, to assess their potential value in additional diagnoses.

Among the 519 ClinVarome only genes in December 2019, 15 genes were in the first cluster, 60 genes were in the second cluster (ie, 75 high-confidence genes), 140 genes were in the third cluster, and 304 genes were in the fourth cluster. Then, we monitored their inclusion in the OMIM morbid list in the upcoming months. Among the 519 genes exclusively referenced in ClinVarome in December 2019, 55 were reported OMIM morbid 8 months later in August 2020, including 15 of the 75 (20%) initial high-confidence genes. Moreover, 125 of the 140 OMIM morbid genes additional entries between December 2019 and August 2020 were also referenced in ClinVarome release of August 2020. This observation suggested that candidate genes in

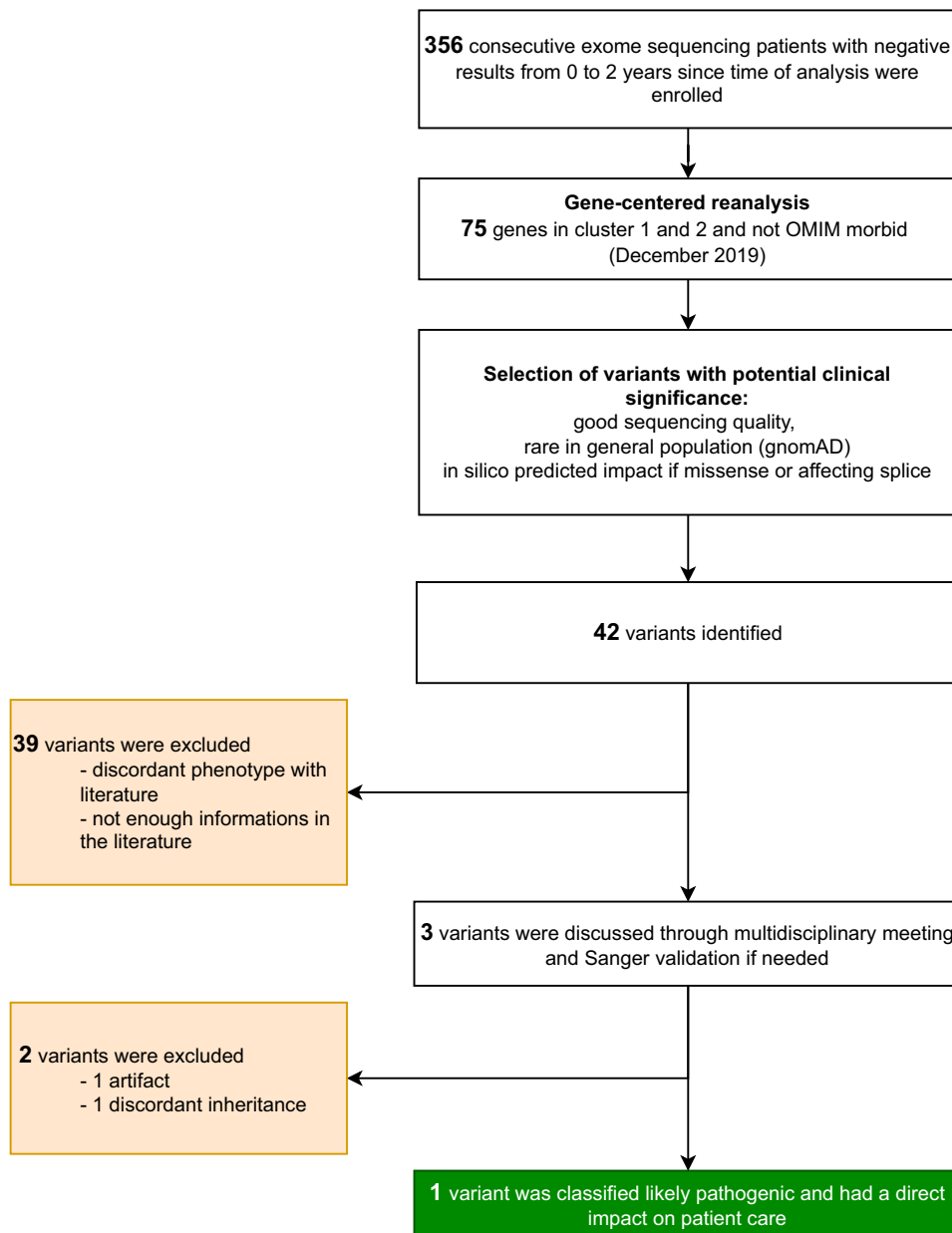
ClinVarome may be considered as diagnostic genes before the OMIM validation of the gene–disease causality.

### Clinical impact of ClinVarome morbid genes not available in OMIM

We evaluated the relevance of this approach by performing a selective reanalysis of a subsample of the new entries in the ClinVarome, focusing only on the 75 genes that were absent from OMIM morbid list and were referenced in ClinVarome’s first and second clusters (gene-centered reanalysis). This experiment highlighted 42 variants in 356

negative exome sequencing data. In this data set, 42 variants were prioritized and were proposed for further interpretation. Among them, 39 were excluded by the expert. The experts’ arguments included the presence of variants unrelated to the disease phenotype or a single case series available in the literature. A total of 3 variants were further explored with Sanger sequencing validation, of which 2 were excluded because of artifact status or discordant inheritance pattern (Figure 5).

Overall, this method could ascertain a new diagnosis from the 356 negative exome sequencing data. A nonsense *DLG4* variant NM\_001128827.1:c.1840C>T was reported



**Figure 5** Experimental design for a targeted gene-centered reanalysis. These 75 genes were reported in ClinVarome and not in OMIM and classified as related to a disease (clusters 1 and 2). This list of 75 genes was used for the reinterpretation of negative exome sequencing data ( $n = 346$ ). Green box represents new diagnosis. Orange boxes represent excluded variants. gnomAD, Genome Aggregation Database.



as likely pathogenic, responsible for the patient's phenotype (intellectual disability and microcephaly). Although the first report of *DLG4* association to intellectual developmental disorder was described back in 2016, this gene–disease association was added to the OMIM database only in February 2020.

### ClinVarome comparison with the GenCC database

A comparison of gene–disease validity confidence and exhaustivity of ClinVarome with the GenCC database was performed. In October 2021, there was a 65% (3332 of 5187) gene overlap between the 2 databases. Nonoverlapping genes represent mostly the uncertain gene–disease associations from these 2 databases. Exclusive genes in GenCC ( $n = 334$ ) were significantly enriched in orange and red genes (151 of 745 orange genes [ $P < .0001$ ], 158 of 252 red genes [ $P < .0001$ ]). Exclusive genes in ClinVarome ( $n = 1471$ ) were significantly enriched in third and fourth cluster genes (407 of 501 third cluster genes [ $P < .0001$ ], 448 of 743 fourth cluster genes [ $P < .0001$ ]). The 2 databases present a high concordance in gene–disease association confidence (Supplemental Table 8).

### Discussion

With the increasing amount of genetic testing performed in health care, there is a critical need for standardized methods to enable prospective genomic data reinterpretation in clinical routine. Through the reassessment of variant pathogenicity and gene–phenotype associations in ClinVar, Genome Alert!'s data mining method proposes the automatic report of a handful of variants that can reasonably be manually interpreted. Our method was applied to a multicentric series of 4929 sequencing tests with various local bioinformatic systems. Genome Alert! successfully allowed new diagnoses in targeted and exome sequencing through query of laboratory's VCFs or variant database and proposed a portable and open-source framework for an automated reanalysis of sequencing data.

Retrospective monitoring of the cutting-edge medical literature on existing genomic data is a major concern for paving the way to genomic medicine.<sup>30</sup> There are numerous technical and medical challenges in setting up a routine procedure for reanalysis. This work explored the dynamics of change across all fields of genomic medicine in ClinVar.

Several medical indications for genomic testing were noticed to bear numerous changes in variant classification. Retrospective analysis of the ClinVar database provided an estimation of new clinically relevant information reported each month, which may lead to additional diagnoses in the existing data.<sup>31</sup> Overall, 9.94 % (1125) of likely pathogenic variants were eventually downgraded and reclassified as benign variants, likely benign variants, variants of uncertain significance, or variants with conflicting interpretation of

pathogenicity in ClinVar over the study period (Supplemental Table 3). This analysis highlights the required carefulness in returning results to the families for likely pathogenic variants because such information could be used for genetic counseling and patient management.

Genome Alert! methods are based on the processing of submissions from the ClinVar full XML release, with no distinction made between submissions with different contexts (eg, somatic or germline status and distinct conditions). Besides, Genome Alert! attributes a unique variant ID on the basis of VCF nomenclature. As such, these variants with potential clinical significance reported by Genome Alert! should be queryable a priori in each genomic center. However, VCF nomenclature is not easy to use with complex variation, which could lead to errors. A switch to the Variation Representation specification from the Global Alliance for Genomics and Health could provide an interesting improvement step.

Clinical effect of changes in variant classification (variant-centered reanalysis) provided in our targeted and exome sequencing cohort provided an additional diagnosis per 1000 analyses. Because time from initial analysis varies from 0 to 2 years, this diagnostic yield will certainly increase with time. This automated system is better for large cohorts of targeted sequencing, with a low number of variants to reinterpret and reaching 10% diagnostic yield in the re-examined variants. Recent literature emphasizes the importance of a standardized procedure adapted for sequencing data reanalysis for considering few candidate variants after an accurate annotation of new gene–phenotype associations and filtering procedure.<sup>30</sup>

A particular effort was made to evaluate confidence in the reported information to reach a consensus across multiple annotations. The prospective reassessment of ClinVar highlighted numerous conflicts in variant classification. Although our system rarely reclassifies variants with conflicting interpretations, this automatic reclassification method aims to at least remove these potential errors. The expert review of ClinVCF automatic reclassification validates this method on the basis of outlier submission removal using the IQR method, and succeeds in reclassifying abnormalities such as the *HFE* pathogenic variant NM\_000410.3:c.845G>A. This work highlights the value of the persistence over time of a classification for relevant genomic information. This work specifically focused on oncogenetics and cardiogenetics, fields in which variant interpretations are particularly conflicting and shifting.<sup>32,33</sup> Overall, in the ClinVar database, 188 variants could be reclassified in 29 months (ranging from 2017 to 2019). After 8 months, in August 2020, a total of 307 variants were reclassified, highlighting the importance of a systematic and partially automated variant reassessment (Supplemental Figure 2).

Existing literature for gene-centered reanalysis has emphasized the importance of OMIM as an updated resource but not exhaustive.<sup>34</sup> To explore and evaluate specifically the ClinVar database for gene-centered reanalysis, we chose to focus our reanalysis on 75 high-confidence ClinVarome

morbid genes (first and second clusters) not available in OMIM morbid genes list. Complementary to OMIM morbid genes, these high-confidence ClinVarome morbid genes from the first and second clusters could provide additional diagnoses in exome or genome sequencing analysis (gene-centered reanalysis). One additional diagnosis was identified with this tight subsampling of variants among the 356 negative exomes, validating the proof of concept. Additional experiments could be performed to fully evaluate the ClinVarome, such as reanalysis with the full list of ClinVarome morbid genes not found in OMIM, additional cohorts, or an extended analysis considering the variants with different phenotypes not reported in the literature.

On the basis of literature data and feature engineering processes from all ClinVarome features during clustering model development, we identified 4 discriminative features for gene–disease clinical validity available in ClinVarome data. Overall, the evaluation relies mainly on the amount of knowledge but also on reported review confidence and more importantly on the time-scale of entries. The Genome Alert! gene-curation via machine learning methods provides an original attempt for automated evaluation of gene confidence in disease. Genome Alert! proposes a standardized clinical validity confidence score that could allow a prospective gene–phenotype association assessment. As such, this approach could be useful to update in silico gene panels. This procedure proposes a complementary approach to the aggregation of multiple expert-reviewed databases such as DDG2P, Genomic England PanelApp, or ClinGen gene–disease validity available in the GenCC database.<sup>35</sup> However, ClinVarome gene–disease validity confidence is defined for all diseases associated with a gene, which is less precise than curations submitted to the GenCC database. As ClinVarome is a more exhaustive database, this resource could prioritize genes to be curated by GenCC submitters, particularly in the first and second clusters.

In summary, Genome Alert! highlights changes with potential clinical significance and provides a large retrospective study of a partially automated system for sequencing data reinterpretation. This procedure enables the systematic and reproducible reinterpretation of acquired sequencing data in a clinical routine, with a limited human resource effect and a diagnostic yield improvement. Genome Alert! provides an open-source accessible framework to the community, thus hoping to be applicable in every genetic center.

## Data Availability

Software summary

Project name: Genome Alert!

Project home page: <https://genomealert.univ-grenoble-alpes.fr/>

Operating system(s): UNIX (Mac, Linux)

Programming language: Nim, Python, R

License: Apache Licence 2.0

Any restrictions to use by nonacademics: No

Genome Alert! results are publicly available at <https://genomealert.univ-grenoble-alpes.fr/>. Relevant data used to generate Genome Alert! results are available from ClinVar FTP (all monthly ClinVar full XML release data were downloaded from <https://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/>) and in the following resources: OMIM (<https://omim.org/>), Genomic England PanelApp (<https://panelapp.genomicsengland.co.uk/>), and RefSeq annotation ([ftp://ftp.ncbi.nlm.nih.gov/refseq/H\\_sapiens/annotation/GRCh38\\_latest/refseq\\_identifiers/GRCh38\\_latest\\_genomic.gff.gz](ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz)). All codes for generating Genome Alert! procedures are available at public GitHub repositories: ClinVCF tool for ClinVar XML full release processing and extraction to VCF format (<https://github.com/SeqOne/clinvcf>), Variant Alert! tool to compare ClinVCF release ([https://github.com/SeqOne/variant\\_alert](https://github.com/SeqOne/variant_alert)), ClinVarome tool to evaluate clinical validity of ClinVar morbid genes (<https://github.com/SeqOne/clinvarome>), and the Genome Alert! shiny app ([https://github.com/SeqOne/GenomeAlert\\_app](https://github.com/SeqOne/GenomeAlert_app)).

## Acknowledgments

We sincerely thank all patients, clinicians, biologists, and bioinformaticians involved in this project. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

## Author Information

Conceptualization: K.Y., F.L., S.Ca., D.B., A.-L.B., J.A., G.N., J.T., N.P.; Data Curation: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R.; Formal Analysis: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.C., Q.F., J.A.; Funding Acquisition: J.T., N.P.; Methodology: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.Ca., Q.F., J.A.; Project Administration: A.-L.B., J.A., G.N., J.T., N.P.; Resources: K.Y., S.Co., M.B., R.L., Q.F., A.D., N.S., S.B., J.A.; Software: K.Y., S.Co., M.B., R.L., Q.F., A.D., N.S., S.B., J.A.; Supervision: A.-L.B., J.A., G.N., J.T., N.P.; Validation: J.A., G.N., J.T., N.P.; Visualization: A.-L.B., D.B., J.A., D.G., L.M., G.N., J.T., N.P.; Writing-original draft: K.Y., F.L., Q.F., Q.T., J.-M.H., D.B., G.N., J.T., N.P.; Writing-review and editing: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.Co., M.B., R.L., Q.F., S.Ca., Q.T., A.D., N.S., J.-M.H., N.D.-F., A.-L.B., S.B., D.B., J.A., D.G., L.M., G.N., J.T., N.P.

## Ethics Declaration

Patients referred to the Eurofins Biomnis laboratory, Cerba laboratory, and CHU de Rouen Molecular Genetics laboratory

provided written consent for analysis of their DNA using next-generation sequencing, including research analysis for the purpose of obtaining a molecular diagnosis. Sequencing samples were de-identified. Local Ethics Committee of the CHU Grenoble-Alpes approved the study. Patients or legal guardians provided informed written consent for genetic analyses in a medical setting. This research conforms to the principles of the Helsinki Declaration.

## Conflict of Interest

K.Y., M.B., R.L., Q.F., A.D., N.S., D.B., A.-L.B., and N.D.-F. are partially or fully employed by SeqOne Genomics; J.M.-H., S.B., J.A., and N.P. hold shares in SeqOne Genomics; D.T., A.B., A.L., and J.-M.C. are partially or fully employed by Laboratoire Cerba. V.G. and L.R. are partially or fully employed by Laboratoire Eurofins Biomnis. All other authors declare no conflicts of interest.

## Additional Information

The online version of this article (<https://doi.org/10.1016/j.gim.2022.02.008>) contains supplementary material, which is available to authorized users.

## Affiliations

<sup>1</sup>Institute for Advanced Biosciences, Centre de recherche UGA / Inserm U 1209 / CNRS UMR 5309, Grenoble, France; <sup>2</sup>SeqOne Genomics, Montpellier, France; <sup>3</sup>Department of Genetics and Reference Center for Developmental Disorders, Normandy Center for Genomic and Personalized Medicine, Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, F 76000, Rouen, France; <sup>4</sup>Laboratoire Cerba, Saint-Ouen-l'Aumône, France; <sup>5</sup>Laboratoire Eurofins Biomnis, Lyon, France; <sup>6</sup>Unité Fonctionnelle de Cardiogénétique et Myogénétique, Centre de Génétique, Hôpitaux Universitaire Pitié Salpêtrière-Charles Foix, Paris, France; <sup>7</sup>Grenoble Institut Neurosciences, GIN, Inserm U1216, Université de Grenoble Alpes, Grenoble, France; <sup>8</sup>Medical Genetic Department for Rare Diseases and Personalized Medicine, Montpellier University Hospital, Montpellier, France; <sup>9</sup>Soins Intensifs Néphrologiques et Rein Aigu, Hôpital Tenon, Assistance Publique des Hôpitaux de Paris, Paris, France; <sup>10</sup>UMR\_S1155, INSERM, Sorbonne Université, Paris, France

## References

- Adams DR, Eng CM. Next-generation sequencing to diagnose suspected genetic disorders. *N Engl J Med*. 2018;379(14):1353–1362. <http://doi.org/10.1056/NEJMra1711801>.
- Shendure J, Findlay GM, Snyder MW. Genomic medicine—progress, pitfalls, and promise. *Cell*. 2019;177(1):45–57. <http://doi.org/10.1016/j.cell.2019.02.003>.
- Dollfus H. Le plan France Médecine Génomique 2025 et les maladies rares. *Med Sci (Paris)*. 2018;34(Hors série n°1):39–41. <http://doi.org/10.1051/medsci/201834s121>.
- Turro E, Astle WJ, Megy K, et al. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*. 2020;583(7814):96–102. <http://doi.org/10.1038/s41586-020-2434-2>.
- Richards S, Aziz N, Bale S, 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):405–424. <http://doi.org/10.1038/gim.2015.30>.
- Nykamp K, Anderson M, Powers M, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med*. 2017;19(10):1105–1117. Published correction appears in *Genet Med*. 2020;22(1):240–242. <https://doi.org/10.1038/gim.2017.37>.
- Tavtigian SV, Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*. 2018;20(9):1054–1060. <http://doi.org/10.1038/gim.2017.210>.
- Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. *Bioinformatics*. 2019;35(11):1978–1980. <http://doi.org/10.1093/bioinformatics/bty897>.
- Nambot S, Thevenon J, Kuentz P, et al. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genet Med*. 2018;20(6):645–654. <http://doi.org/10.1038/gim.2017.162>.
- Wright CF, McRae JF, Clayton S, 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*. 2018;20(10):1216–1223. <http://doi.org/10.1038/gim.2017.246>.
- Bombard Y, Brothers KB, Fitzgerald-Butt S, et al. The responsibility to recontact research participants after reinterpretation of genetic and genomic research results. *Am J Hum Genet*. 2019;104(4):578–595. <http://doi.org/10.1016/j.ajhg.2019.02.025>.
- Clayton EW, Appelbaum PS, Chung WK, Marchant GE, Roberts JL, Evans BJ. Does the law require reinterpretation and return of revised genomic results? *Genet Med*. 2021;23(5):833–836. <http://doi.org/10.1038/s41436-020-01065-x>.
- Carrieri D, Howard HC, Benjamin C, et al. Recontacting patients in clinical genetics services: recommendations of the European Society of Human Genetics. *Eur J Hum Genet*. 2019;27(2):169–182. <http://doi.org/10.1038/s41431-018-0285-1>.
- Deignan JL, Chung WK, Kearney HM, et al. Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2019;21(6):1267–1270. <http://doi.org/10.1038/s41436-019-0478-1>.
- Liu P, Meng L, Normand EA, et al. Reanalysis of clinical exome sequencing data. *N Engl J Med*. 2019;380(25):2478–2480. <http://doi.org/10.1056/NEJMc1812033>.
- James KN, Clark MM, Camp B, et al. Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. *NPJ Genom Med*. 2020;5:33. <http://doi.org/10.1038/s41525-020-00140-1>.
- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(Database issue):D789–D798. <http://doi.org/10.1093/nar/gku1205>.
- Rehm HL, Berg JS, Brooks LD, et al. ClinGen—the clinical genome resource. *N Engl J Med*. 2015;372(23):2235–2242. <http://doi.org/10.1056/NEJMs1406261>.
- Philippakis AA, Azzariti DR, Beltran S, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat*. 2015;36(10):915–921. <http://doi.org/10.1002/humu.22858>.



20. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42(Database issue):D980–D985. <http://doi.org/10.1093/nar/gkt1113>.
21. Tumienė B, Maver A, Writzl K, et al. Diagnostic exome sequencing of syndromic epilepsy patients in clinical practice. *Clin Genet.* 2018;93(5):1057–1062. <http://doi.org/10.1111/cge.13203>.
22. Pengelly RJ, Ward D, Hunt D, Mattocks C, Ennis S. Comparison of Mendeliome exome capture kits for use in clinical diagnostics. *Sci Rep.* 2020;10(1):3235. <http://doi.org/10.1038/s41598-020-60215-y>.
23. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2011;1(1):73–79. <http://doi.org/10.1002/widm.2>.
24. Thomann A, Halachev M, McLaren W, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat Commun.* 2019;10(1):2373. <http://doi.org/10.1038/s41467-019-10016-3>.
25. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019;51(11):1560–1565. <http://doi.org/10.1038/s41588-019-0528-2>.
26. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434–443. Published correction appears in *Nature.* 2021;590(7846):E53. Published correction appears in *Nature.* 2021;597(7874):E3–E4. <https://doi.org/10.1038/s41586-020-2308-7>.
27. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47(D1):D886–D894. <http://doi.org/10.1093/nar/gky1016>.
28. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* 2020;12(1):103. <http://doi.org/10.1186/s13073-020-00803-9>.
29. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* 2014;42(22):13534–13544. <http://doi.org/10.1093/nar/gku1206>.
30. Matalonga L, Hernandez-Ferrer C, Piscia D, et al. Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. *Eur J Hum Genet.* 2021;29(9):1337–1347. Published correction appears in *Eur J Hum Genet.* 2021;29(9):1466–1469. <https://doi.org/10.1038/s41431-021-00852-7>.
31. Landrum MJ, Kattman BL. ClinVar at five years: delivering on the promise. *Hum Mutat.* 2018;39(11):1623–1630. <http://doi.org/10.1002/humu.23641>.
32. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med.* 2016;375(7):655–665. <http://doi.org/10.1056/NEJMsa1507092>.
33. Li D, Shi Y, Li A, et al. Retrospective reinterpretation and reclassification of BRCA1/2 variants from Chinese population. *Breast Cancer.* 2020;27(6):1158–1167. <http://doi.org/10.1007/s12282-020-01119-7>.
34. Bruel AL, Nambot S, Quéré V, et al. Increased diagnostic and new genes identification outcome using research reanalysis of singleton exome sequencing. *Eur J Hum Genet.* 2019;27(10):1519–1531. <http://doi.org/10.1038/s41431-019-0442-1>.
35. Lazo de la Vega L, Yu W, Machini K, et al. A framework for automated gene selection in genomic applications. *Genet Med.* 2021;23(10):1993–1997. <http://doi.org/10.1038/s41436-021-01213-x>.

## **Supplementary Methods**

Supplementary Methods S1: ClinVCF gene name attribution and classification attribution.

Supplementary Methods S2: Genome Alert! comparison with public database resources

Supplementary Methods S1: ClinVCF gene name attribution and classification attribution.

For each variant entry, ClinVCF query the gene name in a two-step process. When a gene symbol is provided and semantically correct according to the NCBI RefSeq GFF, the gene symbol is retained. Otherwise, the variant entry is annotated by ClinVCF according to the NCBI GFF. Then, ClinVCF proposes a consensus classification according to ClinVar policies (aggregation of ACMG/AMP variant classifications interpretations provided in submitted records per variant, according to the level of expertise of submitters) and gathers additional information provided by submitters (e.g. clinical terms or disease name).

In detail, ClinVCF processes VCV interpretations by gathering MeasureSet tags in the ClinVar XML Full Release. We filter GenotypeSet with heterozygous compound value and keep submissions only with standardized ACMG/AMP classifications (which remove Pharmacogenomics haplotypes). Per VCV, a submitter vote is only counted once (submitter ID check). As with ClinVar aggregation policies, when there is a submission from an Expert panel or from a group providing practice guidelines (such as ClinGen), only the interpretation from that group is reported in the aggregate record, even if other submissions provide different interpretations. If there are submissions with one star or more, we only use these submissions in the aggregation. Otherwise, we use all submissions. No distinction was made between somatic and germline submissions. A comparison between NCBI's ClinVar VCF and ClinVCF VCF is possible via the *compvcf* binary file available in the ClinVCF GitHub repository.

In December 2019, ClinVCF provided almost perfect concordance (99.99%) CLNSIG and REVSTAT (99.9%) tags with the NCBI's ClinVar VCF, except some variations because of non-ACMG standard submissions. 2 variants are missing in ClinVCF VCF (Pharmacogenomics haplotype as expected). ClinVCF VCF has 90 additional variants missed in NCBI's ClinVar VCF; no clear patterns were observed to explain this missing information.

ClinVCF provides a three-tier reclassification confidence score. We reclassify variants from conflicting status to likely pathogenic or pathogenic and likely benign or benign status, with a default first-tier confidence score. To ascertain the robustness of this reclassification method, we

have evaluated this automatic reclassification when adding a variant of unknown significance (VUS) in the data submissions. Adding noise can be considered as a defensive approach, in a similar idea to what is being used in deep learning <sup>1</sup>. This test aims at verifying if the amount of data is sufficient to draw similar conclusions, in the event of an additional virtual VUS submission (second-tier confidence). As some reclassifications only rely on likely pathogenic submissions, a definitive reclassification is performed only if at least one pathogenic or benign submission is available (third-tier confidence). As an output, ClinVCF writes a VCF v4.2 file adding the following annotations if an automatic reclassification is performed: proposed reclassification in CLNSIG, ClinVar conflicting interpretations of pathogenicity stats in OLD\_CLNSIG, reclassification confidence score in CLNRECSTAT.

#### Supplementary Methods S2: Genome Alert! comparison with public database resources

Monthly ClinVar full XML release data from 2017-06-20 to 2019-12-01 and from 2019-12-01 to 2020-08-03 were downloaded. From 2017-06-20 to 2019-12-01 data were used for the ClinVar retrospective and the sequencing analysis reinterpretation. The impact of changes were measured on gene groups based on the *in silico* gene panels and disease groups from the Genomics England PanelApp <sup>2</sup> API on 01-06-2020.

Gene symbols gathered from multiple resources (OMIM, ClinVar, and PanelApp) were unified with their NCBI Gene ID (via NCBI RefSeq annotation). To evaluate the exhaustivity of ClinVar morbid gene knowledge, a comparison between the list of all ClinVar morbid genes named *ClinVarome* with the gold standard OMIM database morbid clinical gene list was performed <sup>3</sup>. An OMIM gene is defined as a morbid clinical gene if at least one phenotype or disease syndrome was associated with the gene at that time.

ClinVar data from 2019-12-01 to 2020-08-03 were used to validate ClinVarome. Identification of the gene-phenotype morbid list was made through the OMIM morbid map list (downloaded on 2019-11-14 and 2020-08-24) via the OMIM API<sup>3</sup>.

GenCC data (<https://thegencc.org/>) were downloaded in 2021-10-29 and were compared with the October 2021 release of ClinVarome. GenCC submissions were summarized into 3 categories

(Green, Orange, Red). Green corresponds to Strong or Definitive classification. Orange corresponds to Moderate, Supportive, and Limited classification. Red corresponds to Disputed, Refuted evidence, Animal model, and No known disease relationship classification. If there are only concordant submissions, then the gene category is corresponding to the category submission. If there are Orange and Red or Green submissions, then the gene status will be Red or Green. Conflicting submissions may be observed when Green and Red submissions are available for a gene and were removed from comparative analysis. Enrichment of genes was performed by the scipy Exact Fisher test (one-sided, alternative “greater”) with Bonferroni correction.

#### References:

1. Zheng S, Song Y, Leung T, Goodfellow I. Improving the Robustness of Deep Neural Networks via Stability Training. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Published online 2016. doi:10.1109/cvpr.2016.485
2. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019;51(11):1560-1565.
3. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 2015;43(Database issue):D789-D798

## **Supplemental Tables**

Supplementary Table S1. Genome Alert! Clinical impact change status definition.

Supplementary Table S2. Monitoring of variant classification specific for Genomics England PanelApp panels from July 2017 to December 2019.

Supplementary Table S3. Variant classification evolution in December 2019 among 144.943 variants available in June 2017.

Supplementary Table S4. List of reviewed variants by clinical experts for validation of Genome Alert! automatic reclassification of conflicting interpretation of pathogenicity status.

Supplementary Table S5. Genome Alert! variant classification tracking system analysis of laboratory 1 variant database from the oncogenetic targeted sequencing panel.

Supplementary Table S6. Genome Alert! variant classification tracking system analysis of oncogenetic panels from laboratory 2.

Supplementary Table S7. Median per cluster of the four variables used to classify gene-disease clinical validity.

Supplementary Table S8. Comparison between ClinVarome and GenCC gene-disease association confidence in November 2021.

Supplementary Table S1. Genome Alert! Clinical impact change status definition.

Previous	New	Clinical impact
<b>Variant</b>		
Absent	..	null
	Benign and/or likely_benign	null
	Uncertain_significance or Conflicting_interpretations_of_pathogenicity	unknown
	Likely_pathogenic	major: new pathogenicity
	Pathogenic	major: new pathogenicity
Benign and/or likely_benign	..	deleted from database
	Uncertain_significance or Conflicting_interpretations_of_pathogenicity	unknown
	Likely_pathogenic	major: new pathogenicity
	Pathogenic	major: new pathogenicity
Uncertain_significance or Conflicting_interpretations_of_pathogenicity	..	deleted from database
	Benign and/or likely_benign	minor
	Likely_pathogenic	major : new pathogenicity
	Pathogenic	major : new pathogenicity
Likely_pathogenic or Pathogenic/Likely_pathogenic	..	deleted from database
	Benign and/or likely_benign	major: revoked pathogenicity
	Uncertain_significance or Conflicting_interpretations_of_pathogenicity	major : revoked pathogenicity
	Pathogenic	minor
Pathogenic	..	deleted from database
	Benign and/or likely_benign	major: revoked pathogenicity
	Uncertain_significance or Conflicting_interpretations_of_pathogenicity	major: revoked pathogenicity
	Likely_pathogenic	minor
<b>Gene</b>		
..	NEW_PATHOGENICITY	major : new pathogenicity
Any	LOST_PATHOGENICITY	major: revoked pathogenicity

Supplementary Table S2. Monitoring of variant classification specific for Genomics England PanelApp panels from July 2017 to December 2019.

In Supplementary Excel spreadsheets.



Supplementary Table S3. Variant classification evolution in December 2019 among 144.943 variants available in July 2017. Perc. change category = percentage of variant classification changes by ACMG/AMP variant classifications.

Old classification	type	count	Initial number	Number of changes	Percentage changes	Perc. change category
<i>Benign</i>	warning	59	14255	1502	10.54	0.41
	Likely benign	1262				8.85
	Uncertain significance	180				1.26
	Likely pathogenic	0				0.00
	Pathogenic	1				0.01
<i>Likely benign and Benign/Likely benign</i>	warning	38	31612	2658	8.41	0.12
	Benign	578				1.83
	Uncertain significance	2038				6.45
	Likely pathogenic	2				0.01
	Pathogenic	2				0.01
<i>Uncertain significance and Conflicting interpretations of pathogenicity</i>	warning	555	59683	2157	3.61	0.93
	Benign	185				0.31
	Likely benign	962				1.61
	Likely pathogenic	350				0.59
	Pathogenic	105				0.18
<i>Likely pathogenic and Pathogenic/Likely pathogenic</i>	warning	23	11417	1513	13.25	0.20
	Benign	1				0.01
	Likely benign	2				0.02
	Uncertain significance	1109				9.71
	Pathogenic	378				3.31
<i>Pathogenic</i>	warning	246	27976	2429	8.68	0.88
	Benign	9				0.03
	Likely benign	5				0.02
	Uncertain significance	574				2.05
	Likely pathogenic	1595				5.70

Supplementary Table S4. List of reviewed variants by clinical experts for validation of Genome Alert! automatic reclassification of conflicting interpretation of pathogenicity status.

Conflicting variants reclassified as (likely) pathogenic - December 2019	Gene name	Reclassification confidence	Expert classification
NM_170707.4:c.725C>T	LMNA	1	Class 5
NM_000258.2:c.427G>A	MYL3	1	Class 4
NM_001080116.1:c.494C>T	LDB3	1	Class 4
NC_000011.9(NM_000256.3):c.3815-1G>A	MYBPC3	1	Class 5
NC_000011.9(NM_000256.3):c.2905+1G>A	MYBPC3	1	Class 5
NM_000256.3:c.1504C>T	MYBPC3	3	Class 3 or 4
NM_000257.4:c.2011C>T	MYH7	1	Class 5
NM_000257.4:c.1324C>T	MYH7	1	Class 5
NM_000257.4:c.728G>A	MYH7	1	Class 5
NM_001018005.2:c.644C>T	TPM1	1	Class 4
NC_000018.9(NM_024422.4):c.2125+1del	DSC2	1	Class 5
NC_000001.10(NM_001048174.1):c.849+3A>C	MUTYH	3	Class 5
NM_001128425.1:c.820C>T	MUTYH	1	Class 3
NM_001048171.1:c.267G>A	MUTYH	3	Class 4
NM_000179.2:c.1109T>C	MSH6	1	Class 3
NM_000179.2:c.1295T>C	MSH6	2	Class 3
NM_000179.2:c.3725G>A	MSH6	2	Class 4
NM_000535.7:c.2521del	PMS2	1	Class 4
NM_001126112.2:c.646G>A	TP53	1	Class 5
NM_001126112.2:c.374C>T	TP53	2	Class 5

Supplementary Table S5. Genome Alert! variant classification tracking system analysis of laboratory 1 variant database from the oncogenetic targeted sequencing panel.

In Supplementary spreadsheets.

Supplementary Table S6. Genome Alert! variant classification tracking system analysis of oncogenetic panels from laboratory 2.

In Supplementary spreadsheets.

Supplementary Table S7. Median per cluster of the four variables used to classify gene-disease clinical validity.

<b>cluster name</b>	<b>Highest ACMG/AMP variant classifications found in a variant per gene</b>	<b>Highest ClinVar review confidence (from 0 to 4 stars) found in likely pathogenic or pathogenic variants per gene</b>	<b>Time interval between the first entry and last entry of a likely pathogenic or pathogenic variant per gene (in months)</b>	<b>Count of likely pathogenic or pathogenic variants per gene</b>
4th	4	1	0	1
3rd	5	0	0	2
2nd	5	1	19	7
1st	5	2	26	24

Supplementary Table S8. Comparison between ClinVarome and GenCC gene-disease association confidence in November 2021. Genes in 1st cluster (n=1710) and 2nd cluster (n=1226) are mostly Green (93%, 72% respectively), in comparison of genes in 3rd (n=70) and 4th cluster (n=291) which are mostly Orange and Red (71% and 60%, respectively).

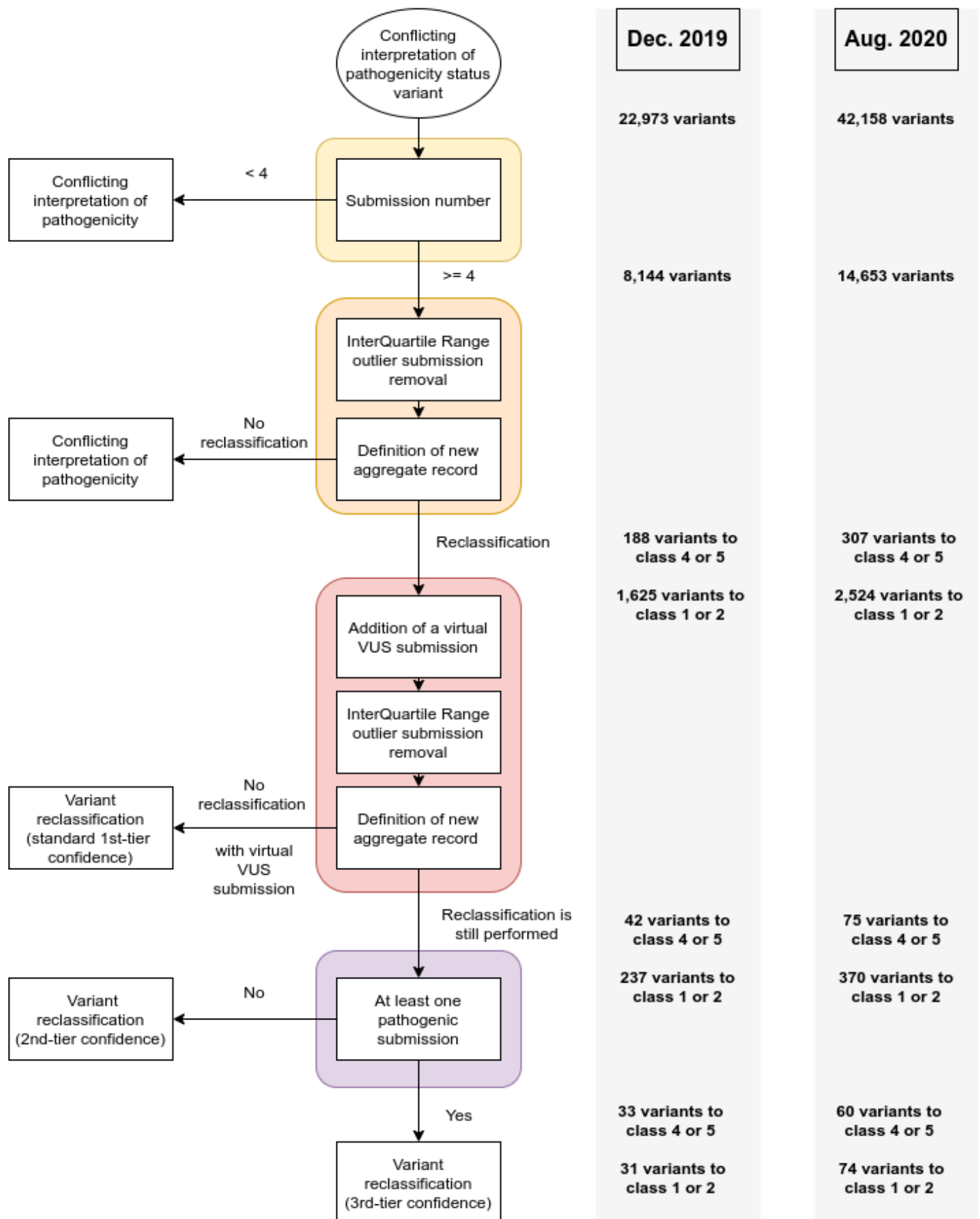
ClinVarome \ GenCC	Green	Orange	Red
<b>1st cluster</b>	1588 (93%)	116 (7%)	6 (<0%)
<b>2nd cluster</b>	886 (72%)	294 (24%)	46 (4%)
<b>3th cluster</b>	26 (29%)	36 (51%)	14 (20%)
<b>4th cluster</b>	115 (40%)	148 (50%)	28 (10%)

## **Supplemental Figures**

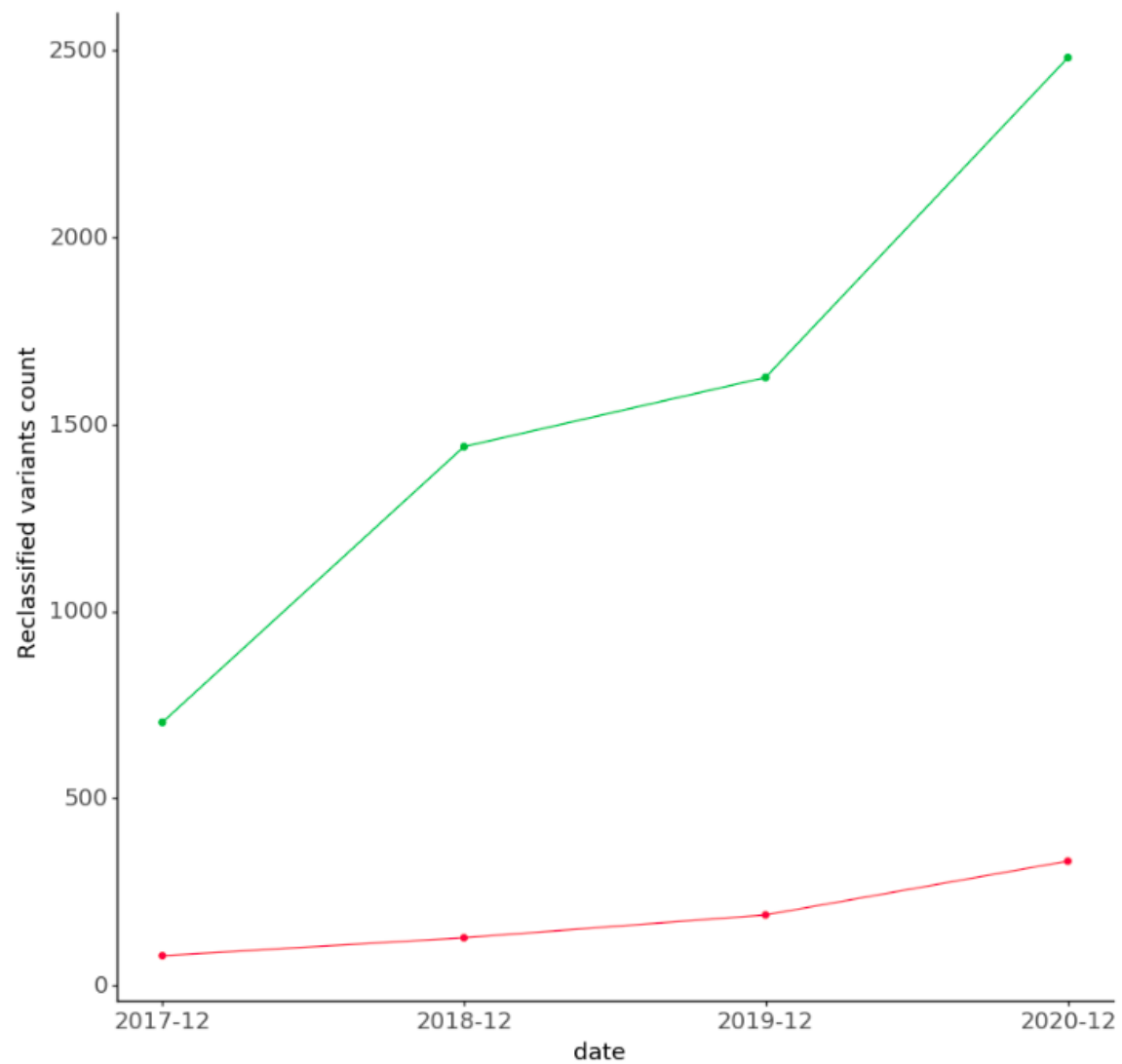
Supplementary Figure S1. Automatic reclassification of variants with conflicting interpretations of pathogenicity workflow from July 2017 to December 2019 and August 2020

Supplementary Figure S2. Evolution of conflicting interpretation of pathogenicity status variants reclassified by Genome Alert! from December 2017 and December 2020.

Supplementary Figure S1. Automatic reclassification of variants with conflicting interpretations of pathogenicity workflow from July 2017 to December 2019 and August 2020



Supplementary Figure S2. Evolution of conflicting interpretation of pathogenicity status variants reclassified by Genome Alert!. Green: reclassification to Likely Benign or Benign class, Red: reclassification to Likely Pathogenic or Pathogenic class.



## Chapter IV - the phenotyping challenge

### *Ask the PhenoGenius*

The key element in genome sequencing interpretation is to find the variant that causes the patient's disease, also called the symptom-gene or genotype-phenotype correlation. Switching from phenotype-first to genotype-first approach in rare diseases diagnosis improved the diagnostic yield of rare diseases. Still, it provided new challenges, such as computational phenotype analysis, where physicians' knowledge and phenotyping need to be digitalized to successfully interpret the massive amount of detected variants. To do so, humans and machines must speak the same language using a common ontology.



# Ontologies for precision medicine

## *Phenotyping*

If the phenotype is the set of observable characteristics of an individual, phenotyping is the process of describing these characteristics deviating from normal morphology, physiology, and behavior. It is the transcription of a physician's clinical and paraclinical examination <sup>57</sup>.

## *Standardization and Ontology*

One working hypothesis is that the standardization of phenotyping is necessary to help us collectively to recognize diseases that are too rare or subtle to diagnose alone. Clinical geneticists are trained to recognize a few hundred conditions, yet it is estimated that around 6,000 rare diseases exist. Teams have been working on an "ontology of human phenotypes " for over ten years <sup>58</sup>. An ontology groups a set of terms linked together as a tree structure. The further one moves away from the "trunk" of the tree (i.e., deeper in the ontology), the more precise the description is. As an example to describe a symptom, we could start from a broad category (e.g., abnormality of the limbs) to a more and more detailed description (e.g., absence of the nail on the 5th finger).

## *Human Phenotype Ontology*

The Human Phenotype Ontology (HPO) is the most commonly used ontology in phenotyping at present <sup>59</sup>. The tree structure of HPO is ordered according to human

development. Each level of the tree corresponds to a code. For example, *partial agenesis of the corpus callosum* is coded HP:0001274. Physicians will collect these codes corresponding to the patient's clinical picture as accurately as possible to weigh the relevance of a particular variation. The combination of symptoms and clinical signs observed in a patient can be processed by diagnostic support algorithms<sup>60</sup>.

To summarize, genetic tests are interpreted in a detailed clinical context. The clinical description using ontology standards allows the integration of medical information into the analyses. HPO ontology represents the ontology of human development and associated phenotypes.

## A constellation of resources

If HPO is a way to describe patients, we need to link a patient's phenotype to known genetic diseases. Since the 1990s and computer democratizations, clinical geneticists query databases such as London Dysmorphology Database or POSSUM as clinical decision support<sup>61</sup>. These genotype-phenotype databases compiled the literature information on genetic diseases, describing the expected phenotypic elements, their frequency, and the genetic causes involved. These databases make it possible to decide, among other features, on the imputability of a variation on all or part of a patient's clinical picture. I would like to give you three examples that we used in our method :

### *OMIM*

OMIM (Online Mendelian Inheritance in Man), currently maintained by the McKusick-Nathans Institute of Genetic Medicine, stems from the work of Dr. Victor McKUSICK in the 1960s to classify Mendelian diseases and their phenotypic traits<sup>62</sup>. The content is edited by a team of researchers who report content related to important articles on the disease or gene in question. More than 6,000 rare diseases with a known molecular basis are listed, and more than 15,000 diseases remain without a known molecular basis.

### *Orphanet*

Orphanet is a French INSERM initiative that was created just before the year 2000, before becoming European and then international, with more than 40 countries involved (<https://www.orpha.net/>). It has two main objectives: to develop a terminology for rare diseases to be integrated into health information systems and provide access to relevant information on rare diseases, their diagnosis, and management for healthcare professionals and patients.

### *DDG2P*

The gene2phenotype dataset (G2P) is produced and curated by UK consultant clinical geneticists for the Deciphering Developmental Disorders (DDD) study<sup>63</sup>. DDG2P integrates data on genes, variants, and phenotypes related to developmental disorders. It is constructed entirely from published literature and is primarily an

inclusion list to allow targeted filtering of genome-wide data for diagnostic purposes, and it also provides HPO terms associated with genes.

### *Aggregating the literature*

Clinical knowledge on genetic diseases is spread into diverse databases, which makes it challenging to use in clinical routine. These databases follow the constant evolution of scientific and medical literature but are manually curated, which provides an unconstant pace in updates. However, most of these databases are not following the HPO format. To facilitate the query on clinical knowledge, some initiatives integrate literature data by querying databases, automatically interpreting the retrieved texts, and looking for gene names, disease names, and signs referenced in HPO. For example, the Monarch Initiative provides a catalog of symptom-gene associations, and NCBI's MedGen provides pages of gene clinical summary <sup>64</sup>. In addition, as most of the clinical information is available in scientific articles, systems like NCBI's LitVar <sup>65</sup> were developed to identify genes and variants contained in articles.

## Motivation

When I was a young resident in medical genetics, I was scared by the amount of knowledge that I needed to learn. I was afraid of being an unworthy physician, afraid of missing a diagnosis. Later in my residency, I understood that more than the

knowledge I earned, I needed to know how to query the clinical knowledge available in the medical literature. But if public resources can efficiently present what you were looking for, systems that help in diagnostic reflection as Phenomizer <sup>66</sup> weren't efficient enough in the medical setting. This would be one of the main challenges for the next generation of clinical geneticists.

Even if we use the HPO to describe a patient's symptoms, I was surprised by the variety of phenotyping practices from my colleagues and myself. Physicians use scheme-induced reasoning based on our own experience <sup>67</sup>. If I see a young child with hypotonia, I will assume that he could have a delayed age of walking and maybe intellectual disability later on. But the available informatic system couldn't guess what physicians' schemes had in mind. It took me a Ph.D. time of reflection and exploration to attempt the digitalization of medical reasoning and make machines able to understand heterogeneous clinical descriptions from physicians'.

Gene panel selection was the current model used by the medical community to digitalize the physician's phenotype in genomic analysis. Initiatives like PanelApp <sup>68</sup> provide human-made groups of genes related to a common clinical entry point. Even if it was an incomplete way to describe a patient, I noticed that physicians already try to gather clinical assumptions around related genes.

Inspired by the clinical entry point - gene association from gene panels and aiming to model medical inductive reasoning, we developed methods based on the association

of symptoms with the same genetic disorder to overcome this phenotyping heterogeneity.

Overall, we described in a scientific article the first analysis of phenotyping practices in a clinical sequencing setting and the development of symptom interaction models in genetic diseases to provide standardized clinical descriptions and interpretable matches between symptoms and genes. We published this study currently as a preprint in medRxiv <sup>69</sup> (<https://www.medrxiv.org/content/10.1101/2022.07.29.22278181>). We have filed two patent applications based on this work. A webapp to use models in clinical practice is accessible at <https://phenogenius.streamlitapp.com>, and the open source code is on GitHub (<https://github.com/kyauy/PhenoGenius>).

I was the principal investigator of the scientific project. I collected the majority of clinical observations from literature, in addition to the cohort from the *PhenoGenius* consortium gathered by Julien Thevenon and Quentin Testard. With the help of Nicolas Duforet from SeqOne Genomics, I programmed scientific experiments, scripts, and webapp to make these methods accessible to the community. I was supervised by Nicolas Duforet, Denis Bertrand, and Julien Thevenon for the scientific exploration and writing of the manuscript.

## **Learning phenotypic patterns in genetic diseases by symptom interaction modeling**

*Kevin Yauy*<sup>1,2,\*</sup>, *Nicolas Duforet-Frebourg*<sup>2</sup>, *Quentin Testard*<sup>1</sup>, *Sacha Beaumeunier*<sup>2</sup>, *Jerome Audoux*<sup>2</sup>, *Benoit Simard*<sup>3</sup>, *Dimitri Larue*<sup>2</sup>, *Michael G. B. Blum*<sup>2</sup>, *Virginie Bernard*<sup>4</sup>, *David Genevieve*<sup>5</sup>,  
*Denis Bertrand*<sup>2</sup>, *PhenoGenius consortium*, *Nicolas Philippe*<sup>2</sup>, *Julien Thevenon*<sup>1,4,\*</sup>,

<sup>1</sup> Institute of Advanced Biosciences, Centre de recherche UGA, Inserm U 1209, CNRS UMR 5309, Grenoble, France.

<sup>2</sup> SeqOne Genomics, Montpellier, France.

<sup>3</sup> OuestWare, Nantes, France.

<sup>4</sup> Reference center for congenital anomalies, Department of Genetics, Genomics and ART, Grenoble-Alpes University-Hospital, Grenoble, France.

<sup>5</sup> Montpellier University, Inserm U1183, IRMB, Reference center for congenital anomalies, Clinical Genetic Unit, Montpellier University Hospital Center, Montpellier, France.

\* Corresponding authors. Email: [kevin.yauy@univ-grenoble-alpes.fr](mailto:kevin.yauy@univ-grenoble-alpes.fr) and [JThevenon@chu-grenoble.fr](mailto:JThevenon@chu-grenoble.fr)

## Abstract

Observing phenotyping practices from an international cohort of 1,686 cases revealed heterogeneity of phenotype reporting among clinicians. Heterogeneity limited their exploitation for diagnosis as only 43% of symptom-gene associations in the cohort were available in public databases. We developed a symptom interaction model that summarized 16,600 terms into 390 groups of interacting symptoms and detected 3,222,053 novel symptom-gene associations. By learning phenotypic patterns in genetic diseases, symptom interaction modeling handled heterogeneity in phenotyping, to the extent of covering 98% of our cohort's symptom-gene associations. Using these symptom interactions improved the diagnostic performance in gene prioritization by 42% (median rank 80 to 41) compared to the best algorithms. Symptom interaction modeling will provide new discoveries in precision medicine by standardizing clinical descriptions.

### One sentence summary

Learning phenotypic patterns in genetic disease by symptom interaction modeling addresses physicians' heterogeneous phenotype reporting.



Precision medicine relies on patient stratification and recognition of clinically relevant groups to improve diagnosis, prognosis, and medical treatment <sup>1</sup>. Phenotyping allows homogeneous groups of individuals to be constituted, where physicians report characteristics deviating from normal morphology, physiology, and behavior using standardized descriptions in the Human Phenotype Ontology (HPO) <sup>2,3</sup>. Despite a common ontology and abundant clinical data, medical records often lack consistency and comparability between descriptions and practitioners, which is referred to as fuzzy matching in phenotype profiles <sup>4</sup>. This inconsistent phenotyping is a major hurdle to fully exploiting the clinical data contained in medical records. Nevertheless, no studies about phenotyping practices in clinical sequencing are known to have been undertaken until now.

## 1. Phenotyping practices in large cohorts

Through four international studies, including 1,686 patients in total, we collected 2501 different symptoms in HPO format and 849 different disease-causing genes <sup>5-7</sup> (Table S1). Nearly half of the patients in the multi-center cohort had symptoms belonging to the *Abnormality of the nervous system* (HP:0000707) and *Abnormality of the musculoskeletal system* (HP:0033127) classes, illustrating the current focus on those rare disorders in clinical practice <sup>8</sup> (Figure S1). Reflecting the genetic heterogeneity of rare diseases, 538 of 849 genes were declared only once in the cohort and the most frequently mutated gene occurred in less than 2% of cases (*ABCC6*, n=21, Table S2).

We observed heterogeneity in HPO selection terms, as 47% of terms were used only once (Figure 1A, Table S3). The median number of HPO terms per physicians' clinical description varied across observations, ranging from three (Peng *et al.* <sup>7</sup>) to seven (PhenoGenius consortium, Seo *et al.* and Trujillano *et al.* <sup>5,6</sup>) (Figure 1B). The heterogeneity of physicians'

clinical descriptions was also observed for patients with identical genetic diagnoses. For genes involved in diagnosis of more than ten patients, 67 % of symptoms were declared in only one clinical description.

To exclude the possibility that the observed heterogeneity was due to variability in clinical examinations, we next investigated whether heterogeneity in clinical descriptions was reported if physicians phenotyped the same clinical observations. We settled on a prospective experiment where 12 clinical geneticists with various levels of expertise (Table S4) were asked to phenotype three independent clinical reports associated with genetic test prescription, i.e. to convert free text to phenotypes in HPO format. We observed heterogeneity in terms of the number and diversity of symptoms declared per clinical observation (Figure 1C). For instance, two to nine symptoms were declared in clinical descriptions of the Kleefstra syndrome observation with the *EHMT1* pathogenic variant. A total of 29 different terms were provided; 17 of these terms were used by two or more physicians, and none of the terms were mentioned by all 12 physicians.

## 2. Quantifying the overlap of symptoms-gene associations between the retrospective cohort and the medical literature

To assess if the clinical descriptions of our cohort matched available knowledge in the medical literature, we mapped the cohort's 11,526 unique symptom-gene associations to the 734,931 associations available in HPO-structured databases (Orphanet, DDG2P <sup>9</sup>, and the Monarch Initiative or MI <sup>3</sup>). From these databases, only 4,913 associations (43%) matched, meaning that 57% were missing (Figure 2A).

As the clinical descriptions of genetic diseases in medical literature are mainly available in free-text format, we developed a text-mining algorithm based on Elasticsearch to extract symptom-gene associations from free-text data in HPO format. Applied to OMIM <sup>10</sup>, MedGen <sup>11</sup>, and abstracts from PubMed, this text-mining algorithm identified an additional 1,049,522 symptom-gene associations. This approach resulted in a 3.2-fold increase in HPO-structured database associations (Figure 2B).

The text-mining algorithm provided symptom-gene associations where symptoms were significantly deeper in the ontology compared to the HPO-structured databases (median depth 6.7 and 5.2 respectively, Kolmogorov-Smirnov test p-value  $< 10^{-215}$ , Figure S2). This underlines the complementarity of these approaches, as illustrated in Figure 2C where *KMT2D* was associated with *Abnormal morphology of the great vessels* (HP:0030962) in the MI database and *Tetralogy of Fallot* (HP:0001636) in the OMIM database. Reflecting the variability across individuals in selecting an HPO term to summarize a clinical observation, 76% of associations were exclusive to one database. We hypothesized that text-mined symptom-gene associations in the literature were related to associations available in HPO-structured databases. This hypothesis embodies the fuzzy phenotyping concept, providing human-determined alternative wordings of the same information.

To evaluate this hypothesis, for each gene we compared the average distance in the ontology of exclusive symptom-gene associations to the MI database and the text-mined OMIM database, respectively the largest database of each type (Figure 2B). Compared with a random choice of an HPO term, the average distance of the exclusive symptom-gene associations was significantly lower, suggesting these associations are related (Kolmogorov-Smirnov test p-value  $< 10^{-215}$ , Figure 2D).

Although in this exercise the number of symptom-gene associations increased from 734,931 (MI, DDG2P, Orphanet database) to 1,784,453 (with associations found with the text mining algorithm), a match with the cohort's symptom-gene associations was only available for 6,226 of 11,526 (57%) associations, meaning that 43% of matches were still missing (Figure 2A).

### 3. From symptom-gene to symptom-symptom associations modeling

We investigated whether modeling associations between symptoms of the same genetic disorder improved matches. As the Human Phenotype Ontology is ordered according to human development, it may not represent the interaction of symptoms in disease (Figure 3A). We explored an alternative approach to measure symptom-symptom associations in genetic diseases. We considered a node similarity algorithm based on a knowledge graph that stored the symptom-gene associations we collected from the literature.

We found a high correlation between symptom-symptom similarity pair scores and their frequency of co-occurrence in clinical observations (Spearman correlation coefficient: 0.99). No correlation was observed between symptom-symptom similarity pair scores and the distance between symptoms in the HPO (Spearman correlation coefficient: -0,02, Figure S3), reflecting that symptom-symptom associations cannot be solely derived from the ontology architecture.

According to similarity score distributions, we posited that similarities above 80% were potential substitutes or highly similar symptoms in diseases (Figure S4). This resulted in the

selection of 565,943 pairs of highly similar symptoms, corresponding to the 10% highest symptom-symptom association scores (Figure 3B). A total of 26% of these pairs were observed for symptoms in the same ontology class (145,611 of 565,943), mostly from the *Abnormality of the musculoskeletal system* (HP:0033127) class (51%, 73,817 of 145,611). Inter-classes pairs of symptoms represented 74% of highly similar symptoms, where the most recurrent pair was *Abnormality of metabolism/homeostasis* (HP:0001939) with *Abnormality of the nervous system* (HP:0000707) (8%, 35,476 of 420,332).

We illustrate these similarities in Figure 3C, using the symptom *Hypotonia* (HP:0001290) reported by six of the 12 practitioners in our exercise on the Kleefstra syndrome with the *EHMT1* pathogenic variant. In the symptom-symptom association graph, the closest term to *Hypotonia* is *Neurodevelopmental delay* (HP:0012758), with a symptom-symptom similarity pair score measuring 86%. In the HPO, these symptoms are separated by ten nodes and belong to two different main classes: *Abnormality of the musculoskeletal system* (HP:0033127) and *Abnormality of the nervous system* (HP:0000707) respectively.

We then investigated to what extent considering two highly similar symptoms as substitutes improved the coverage of symptoms-gene associations. Among the cohort's 11,526 unique symptom-gene associations, only 6,226 associations were found in HPO-structured and text-mined databases, but this number rises to 8,350 when accounting for similarities. Considering substitutes provided additional 1,506,469 symptom-gene associations to the previous 1,784,453 associations from MI, DDG2P, Orphanet, and text-mined databases.

Modeling associations between symptoms revealed a majority of inter-HPO classes included similar symptoms, highlighting the missing aspect of symptom relationships in the HP

ontology. Enhancing symptoms with their highly similar pairs improved coverage of symptom-gene associations in the cohort, but 27% of associations were still missing.

#### 4. From symptom-gene associations to groups of symptoms modeling

Symptom-symptom associations were evaluated independently when identifying substitutes based on node similarity. To gain better coverage of symptom-gene associations, we considered a more elaborate collaborative filtering approach based on non-negative matrix factorization (NMF) <sup>12</sup>.

Using the topic coherence measure <sup>13</sup>, we determined that the 16,660 HPO terms could optimally be reduced to 390 groups of interacting symptoms or phenotypic patterns (Figure S5). Each symptom was positioned in the graph with group weights determined by the algorithm (Figures 4A-4B). Each gene was associated in a median of 36 groups and a group with a median of 501 genes. To compare the recall of the NMF and the node similarity model, we kept only the top 10% of 390 symptom-groups weights (Figure S6). Overall in this selection, there were 43,308 symptom-group associations leading to 5,971,755 pairs of symptoms.

We investigated to what extent the coverage of symptoms-gene associations was improved by considering that two symptoms belonging to the same group were substitutes. Using these pairs of symptom associations enhanced the coverage of symptom-gene associations to 11,340 of the 11,526 associations from the cohort, leaving less than 2% of matches missing. This new manner of detecting associations resulted in the addition of 2,163,663 NMF-based symptom-gene associations to the previous 1,784,453 associations obtained from MI, DD2P,

Orphanet, and text-mined databases. NMF-based symptom-gene associations overlapped with 99% of similarity-based associations (1,497,601 of 1,506,469).

To evaluate if these 390 phenotypic patterns represented the clinical spectrum of genetic diseases, we projected the cohort into the groups of symptoms dimension and performed a UMAP visualization <sup>14</sup>. We applied agglomerative clustering to the cohort and compared clustering patient performance using this projection and the 16,600 HPO dimension. Using the initial list of 16,600 symptoms, 152 patients were found in 14 clusters significantly enriched in symptoms (Fisher exact test with p-value < 0.05 with Benjamini Hochberg correction) (Figures S7-S8). Applying the projection in groups of symptoms, 1,136 patients were found in 51 clusters significantly enriched in groups of symptoms (Figure S9, Figure S10). To evaluate if this projection could standardize clinical descriptions, we applied it to the three clinical reports phenotyped by the 12 physicians in our experiment. We demonstrated the high coherence of our method even with symptom heterogeneity when sufficient numbers of HPO terms were given (*KMT2D report*) (Figure 4C, Figure S11). When fewer than 5 terms were provided, clinical description projections still grouped patients but with lower homogeneity (*EHMT1, C3*).

The delineation of 390-groups of interacting symptoms enabled an increase in coverage of the available knowledge on genetic disorders and provided a way of building on HP ontology to standardize clinical descriptions. Next, we used symptom interaction modeling to develop a phenotype matching system.

## 5. Symptom interaction models as an efficient and robust system for phenotype matching

To evaluate the clinical relevance of symptom interaction models, we designed phenotype matching and diagnostic gene ranking experiments. We defined a phenotype match when at least one symptom in the clinical description was related to the diagnostic gene (Figure 5A). According to the count of matches per gene, a personalized ranked list of genes was provided (Figure S12). These experiments were performed on the clinical observations of 1,686 patients.

Using the HPO-structured databases (MI, DDG2P, Orphanet), we obtained a phenotype match for 1,566 clinical observations with a median diagnostic gene rank of 251. Applying text-mined associations led to a match for 1,628 clinical observations with a median rank of 40 (Figure 5B). The best performance in median diagnostic rank was provided by node similarity symptoms association (median rank 37, compared to 58 with NMF), but NMF was able to get a more exhaustive coverage of clinical observations (1682, compared to 1663 with node similarity). This coverage gap was exclusively observed where the clinical descriptions contained five terms or less (four unmatched descriptions, compared to 25 with node similarity). As each symptom interaction model provides a different level of inductive reasoning, we conditionally applied a model according to the number of symptoms in the clinical description. The combined system, which we called PhenoGenius provided the best performance (median rank 41) and reached a nearly full phenotype match of diagnostic genes (99.8%, 1682/1686) for all clinical descriptions.

To illustrate this phenotype-matching system, we considered a clinical description containing two symptoms of the Kleefstra syndrome observation with the *EHMT1* pathogenic variant:



*Sparse hair* (HP:0008070) and *Moderate global developmental delay* (HP:0011343). There is no match between these terms and *EHMT1* in HPO-structured databases. No match is identified from text-mined symptom-gene associations either. Symptom interaction modeling achieves phenotype matches, ranking 1244 out of 5235 (top 25% of genes) with the similarity model and 851 out of 5235 (top 17% of genes) with the NMF model and PhenoGenius combined system.

We then compared PhenoGenius to four recently published algorithms for phenotype-driven gene prioritization: PhenoApt, Phen2Gene, CADA, and LIRICAL<sup>7,15-17</sup>. Despite using different prioritization methodologies, these four programs demonstrated similar performances in phenotype matching (Figure 5C). Using symptom interaction modeling, PhenoGenius (median rank 41) increased the median diagnostic gene rank by 42% compared to the best competitor, Phen2Gene (median rank 71, 73 to 80 for other methods). This improvement was replicated across each study subgroup in the cohort, highlighting the clinical relevance of symptom interactions in genetic disease models (Figure S13).

To assess the robustness of gene prioritization, we randomly removed each symptom from clinical descriptions with two terms or more and measured the consequence on the disease-causing gene ranking for descriptions in the top-ranked half of the cohort (rank 41 or lower). Overall, 701 clinical descriptions led to 6,331 symptom removal experiments. In most cases, phenotype matching remained robust with symptom removal (Figure 5D). Disease-causing gene ranking was identical in 35% of cases (2,274 of 6,331) and the median of absolute differences between ranks was only one. However, nine extreme drops in the ranking (> 1000) were observed with clinical descriptions with three or fewer terms,

including two complete loss of phenotype matches for descriptions with two symptoms. For clinical descriptions with four or more terms, we found no extreme drops in gene rankings.

## Discussion

This study used symptom interaction modeling to learn phenotyping patterns in genetic diseases. This method adds to the precision medicine toolbox with a way of standardizing clinical descriptions and matching physicians' phenotyping to the medical knowledge of genetic diseases.

This study provides an in-depth analysis of phenotyping clinical practice by analyzing 1,686 phenotyping reports of patients with a definitive genetic diagnosis<sup>5-7</sup>. In addition, a qualitative comparison of three clinical reports phenotyped by 12 physicians was performed. Complementary to recent reports<sup>8,18</sup>, this study provides original insights on heterogeneous patient phenotyping, both in the cohort's clinical descriptions and the medical literature. In our qualitative experiment, the main observation was the diversity of terms chosen by physicians to describe the exact same clinical description. These observations suggest that clinical description should be standardized, following harmonization of symptom description with HP ontology.

As well as encouraging richness of clinical description, tools must address the medical reality of summarized or partial clinical information. Lacking time or omitting symptoms in their clinical routine, physicians provide scanty phenotyping. Symptoms may be chosen based on strong clinical *a priori* or learned phenotypic patterns. Medical inductive knowledge often proposes patterns or groups of hypotheses based on recurrently associated symptoms in the

physician's own experience and in the literature. Defining groups of symptoms represent a natural behavior of medical inductive reasoning<sup>19,20</sup>. This could explain the heterogeneity of phenotyping across clinical observations, independently from the innate clinical heterogeneity of a disorder.

To handle heterogeneous phenotyping, we developed symptom interaction models to standardize clinical descriptions and evaluate their clinical relevance through gene prioritization experiments. Based on symptom interaction models, PhenoGenius decreases the rank of the diagnostic gene by 42% compared to the best competitor. Its simplicity in scoring allows a complete understanding of phenotype matching, thus providing an interpretable measure of potential genotype-phenotype correlation. To lower the risk of missing a phenotype match because of a fuzzy description, clinical descriptions with four or more terms are recommended. Our approach contrasts with state-of-the-art phenotype-driven gene prioritization software, which mostly relies on complex scoring or symptom relationships based on HPO architecture.

Current algorithms address phenotyping heterogeneity using the ontology structure either to extract additional symptom-gene associations from literature or to evaluate the semantic similarity of symptoms<sup>21</sup>. In contrast to these approaches, we used HPO as a dictionary of symptoms and considered relationships between symptoms only through their co-occurrence in genetic diseases found in HPO-structured and text-mined databases. Our algorithm uncovered the missing pieces of medical inductive reasoning in clinical descriptions through symptom similarity modeling and collaborative filtering using NMF methods<sup>12</sup>. As such, projection into the symptoms interaction model dimension could provide a path to standardizing clinical descriptions. Moreover, the application of this algorithm is

reproducible and interpretable, and these features are fundamental in a medical context <sup>22</sup>. In addition, node similarity and NMF allow free association of symptoms, which is important since the same symptom may belong to different disease groups.

Our AI system performed well for gene prioritization. However, evaluation of our system's performance in detecting gene/symptom associations is incomplete. In our international cohort, only 43% of symptom-gene associations were described in public databases. We have shown that the recall rate (percentage of detected associations among known associations) increased when considering similarity measures or techniques based on NMF. However, as the list of associations increased, an increase in recall came at a price of reduced precision, i.e. a reduced proportion of true associations among the detected associations. Evaluation of precision is impossible because some true associations are missing, highlighting the need to improve data sharing of physicians' phenotype information.

As current knowledge overwhelms human learning abilities, an overarching goal in precision medicine is to overcome digital bottlenecks to succeed in deep phenotyping and identification of clinically relevant groups of patients. Progressive adoption of the Monarch Initiative's HPO in clinical symptoms description, the development of automatic extraction of symptoms in HPO format from electronic medical records <sup>23</sup>, and the definition of the Phenopackets standard file format by GAG4H <sup>24</sup> bring the community one step forward. A current challenge is integrating multiple data sources from electronic health records for deep phenotyping <sup>25</sup>. Complementary to this challenge, we seek to standardize and improve the exploitation of clinical descriptions available in clinical practices using symptom interaction models. Long-standing aspirations are to be able to answer the question, "Have I seen a case

like that before?” among extensive clinical data, and to identify undescribed symptom-gene associations <sup>26</sup>.

Clinical description standardization using symptom interaction modeling may overcome several clinical bottlenecks in precision medicine. PhenoGenius is open-source, accessible through an interactive graph browser (<https://github.com/kyauy/PhenoGenius>), and a web app (<https://phenogenius.streamlitapp.com/>). This work paves the way for a set of tools to help identify new genes in disease, expand their clinical spectrum, and provide an easily interpretable clinical decision support system. If we can successfully deal with fuzzy phenotypic profiles and inductive medical reasoning in rare diseases, clinical data can be used for computational phenotype analysis, to improve the feasibility of precision medicine, and to support the adoption of genomic medicine.

### Data availability

The PhenoGenius source code is available for resource generation and scientific experiments in Apache License 2.0, including an interactive graph browser, on GitHub (<https://github.com/kyauy/PhenoGenius>). A web app is accessible at <https://phenogenius.streamlitapp.com>.

### Declaration of Interest

This study has been jointly funded by Association Nationale de la Recherche et de la Technologie (ANRT) and SeqOne Genomics. K.Y., N.D.-F., S.B., J.A., D.L., M.G.B., D.B., and N.P. are partially or fully employed by SeqOne Genomics; D.L., S.B., J.A., and N.P. hold shares in SeqOne Genomics. K.Y., N.D.-F., S.B., D.L., J.A., N.P., and J.T. have filed two patent applications based on this work.

## Contributions

K.Y., N.D.-F., M.G.B., D.B., and J.T. contributed to the writing of the manuscript and generation of figures. K.Y., N.D.-F., Q.T., B.S., V.B., D.B., and J.T. contributed to the analysis of data. K.Y., N.D.-F., Q.T., S.B., J.A., B.S., D.L., N.P., and J.T. developed tools and methods that enabled the scientific discoveries herein. K.Y., N.D.-F., and J.T. contributed to the collection of the PhenoGenius dataset. All authors listed under PhenoGenius Consortium contributed to the generation of the primary data incorporated into the PhenoGenius resource. All authors reviewed the manuscript.

## Acknowledgments

We sincerely thank all patients, clinicians, biologists, bioinformaticians, and data scientists involved in this project. We are grateful to our scientific mentors for their helpful advice: Stanislas Lyonnet and Jean-Louis Mandel. Warm thanks to Jennifer Butt, who provided an extensive English language review of this manuscript. Special thanks are addressed to the Association Francophone de Génétique Clinique (<https://af-gc.fr/>) and to the physicians who performed phenotyping on the three clinical observations: Roseline Caumes, Mélanie Fradin, Anne-Marie Guerrot, Valentin Ruault, Godelieve Morel, Aurelia Jacquette, Gwenaël Le Guyader, Benjamin Dauriat, Geoffroy Delplancq, Bertrand Chesneau, Annick Toutain and Sarah Cluzel. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

**PhenoGenius consortium** : *Yannis Duffourd*<sup>6</sup>, *Nicolas Chatron*<sup>7</sup>, *Cedric Le Maréchal*<sup>8</sup>, *Jean-François Taly*<sup>9</sup>, *Wilfrid Carre*<sup>10</sup>, *Claire Bardel*<sup>11</sup>, *Frederic Tran Mau-Them*<sup>6</sup>, *Marc Planes*<sup>12</sup>, *Marie-Pierre Audrezet*<sup>12</sup>, *Laure Raymond*<sup>9</sup>, *Charles Coutton*<sup>13</sup>, *Pierre Ray*<sup>13</sup>, *Veronique Satre*<sup>13</sup>, *Klaus Dietrich*<sup>13</sup>, *Isabelle Marey*<sup>13</sup>, *Françoise Devillard*<sup>13</sup>, *Radu Harbuz*<sup>13</sup>, *Florence Amblard*<sup>13</sup>,

*Pauline Le Tanno*<sup>13</sup>, *Mouna Barat-Houari*<sup>14</sup>, *Marjolaine Willems*<sup>14</sup>, *Thomas Guignard*<sup>14</sup>, *Sylvie Odent*<sup>15</sup>, *Marie de Tayrac*<sup>10</sup>, *Damien Sanlaville*<sup>7</sup>, *Laurence Faivre*<sup>6</sup>, *Laurent Mesnard*<sup>16</sup>

<sup>6</sup> Inserm UMR 1231 GAD, Genetics of Developmental disorders and Centre de Référence Maladies Rares Anomalies du Développement et syndromes malformatifs FHU TRANSLAD, Université de Bourgogne-Franche Comté, Dijon, France.

<sup>7</sup> Department of Medical Genetics, Lyon University Hospital, 69677 Lyon, France; CNRS UMR 5292, INSERM U1028, CNRL, 69500 Lyon, France; Université Claude Bernard Lyon 1, GHE, 69100 Lyon, France.

<sup>8</sup> Laboratoire de Génétique, UMR 1078 Génétique, Génomique fonctionnelle et Biotechnologies, Inserm, Université de Brest, EFS, CHU Brest, Brest, France.

<sup>9</sup> Service de Génétique, Eurofins Biomnis, Lyon, France

<sup>10</sup> Service de Génétique Moléculaire et Génomique, CHU Rennes, Rennes, France.

<sup>11</sup> Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, Villeurbanne, France. Service de Biostatistique-bioinformatique et plateforme NGS-CHU Lyon, Hospices Civils de Lyon, Lyon, France.

<sup>12</sup> Laboratoire de Génétique, CHU Brest, Brest, France.

<sup>13</sup> Genetic Epigenetic and Therapies of Infertility, Institute for Advanced Biosciences, Inserm U1209, CNRS UMR 5309, Université Grenoble Alpes, 38000, Grenoble, France. CHU de Grenoble, UM de Génétique Chromosomique, 38000, Grenoble, France.

<sup>14</sup> Département de Génétique Médicale, Maladies Rares et Médecine Personnalisée, Univ Montpellier, CHU de Montpellier, CLAD ASOOR Montpellier, France.

<sup>15</sup> Service de Génétique Clinique, Centre Référence "Déficiences Intellectuelles de causes rares" (CRDI), Centre de référence anomalies du développement CLAD-Ouest, CHU Rennes,

35203 Rennes, France; CNRS UMR 6290, Université de Rennes, 2 Avenue du Professeur  
Léon Bernard, 35043 Rennes, France.

<sup>16</sup> Soins Intensifs Néphrologiques et Rein Aigu, Hopital Tenon, Assistance Publique -  
Hopitaux de Paris-Sorbonne Université, Paris, France.

## References

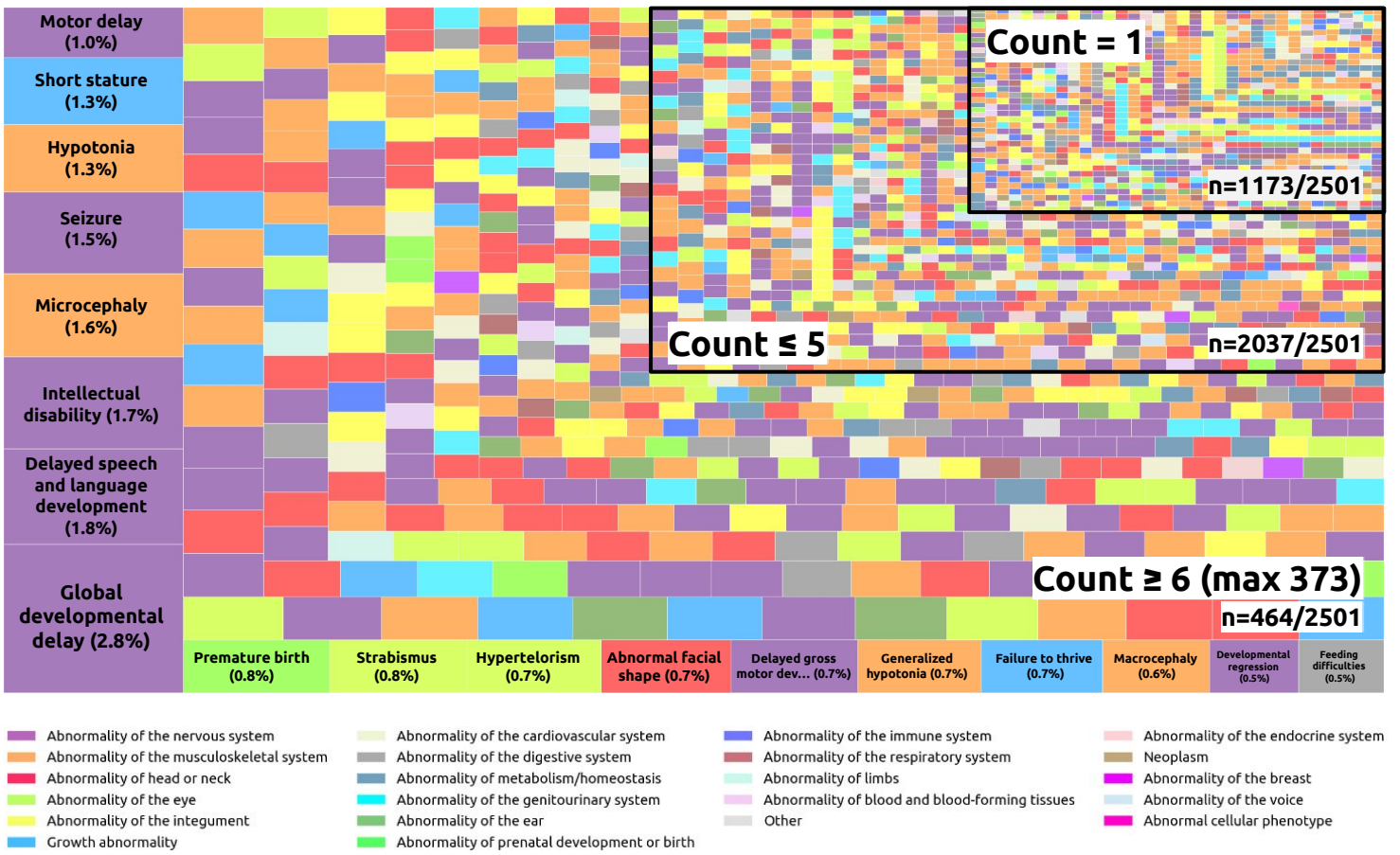
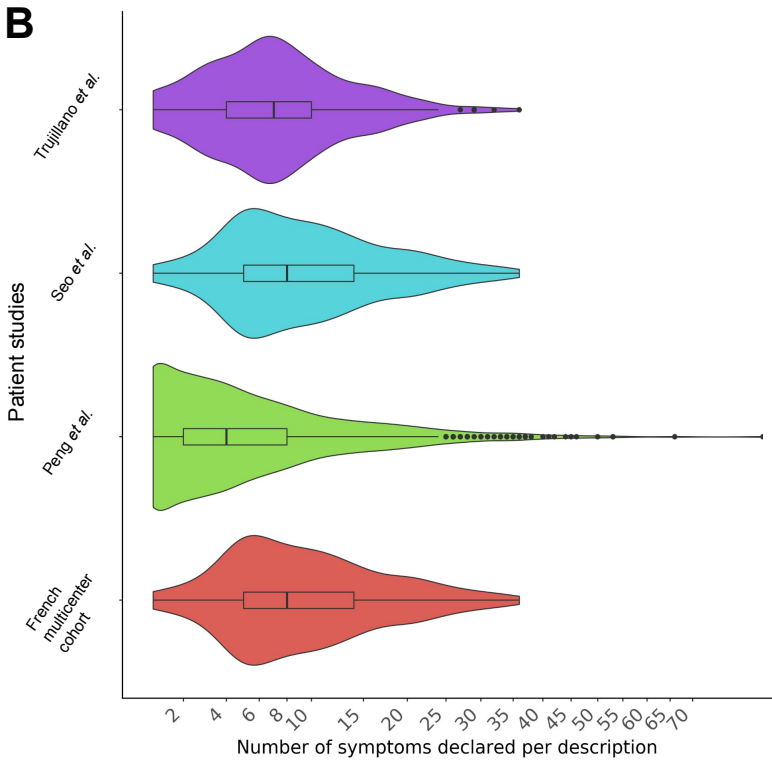
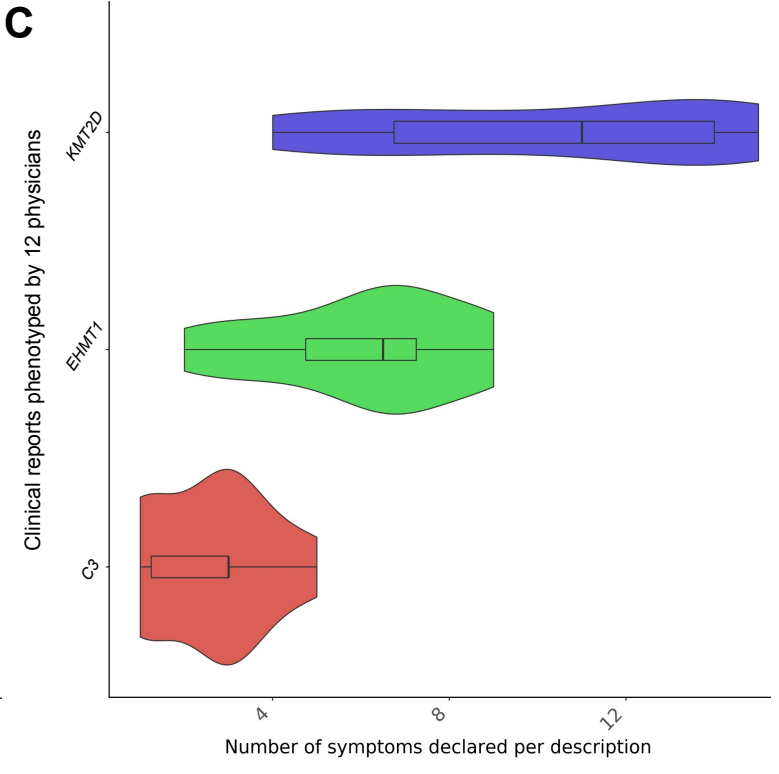
1. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, (2016).
2. Robinson, P. N. Deep phenotyping for precision medicine. *Human Mutation* vol. 33 777–780 Preprint at <https://doi.org/10.1002/humu.22080> (2012).
3. Köhler, S., Kindle, G. & Robinson, P. N. *The Human Phenotype Ontology in 2021*. (2021).
4. Haendel, M. A., Chute, C. G. & Robinson, P. N. Classification, Ontology, and Precision Medicine. *N. Engl. J. Med.* **379**, 1452–1462 (2018).
5. Seo, G. H. *et al.* Diagnostic yield and clinical utility of whole exome sequencing using an automated variant prioritization system, EVIDENCE. *Clin. Genet.* **98**, 562–570 (2020).
6. Trujillano, D. *et al.* Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* **25**, 176–182 (2017).
7. Peng, C. *et al.* CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genom Bioinform* **3**, lqab078 (2021).
8. Osmond, M. *et al.* PhenomeCentral: 7 years of rare disease matchmaking. *Hum. Mutat.* **43**, 674–681 (2022).
9. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using



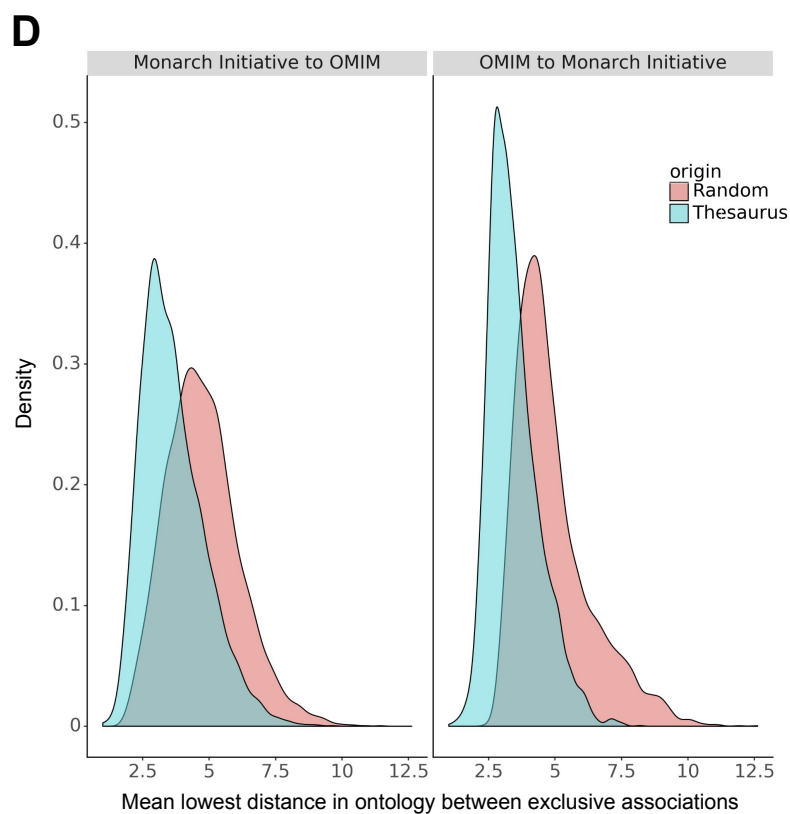
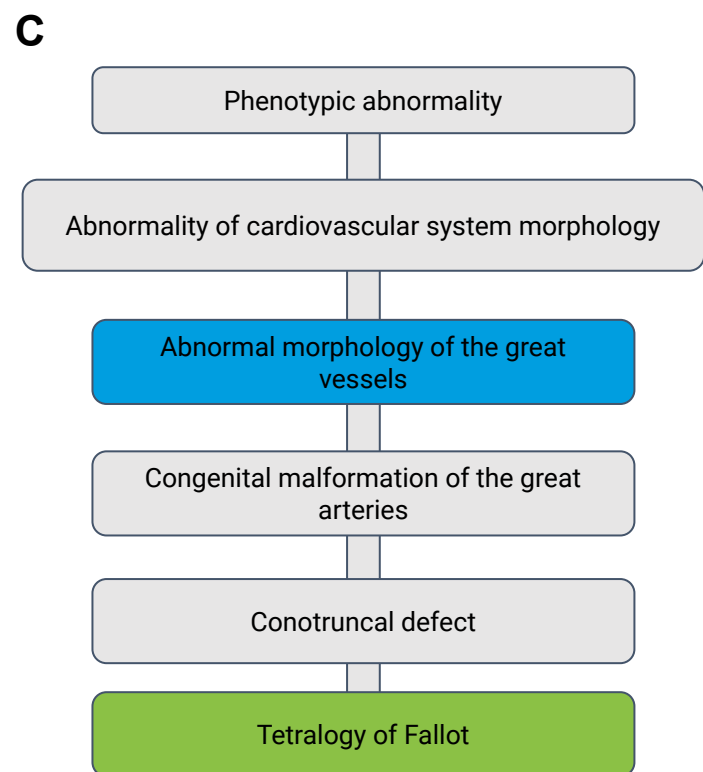
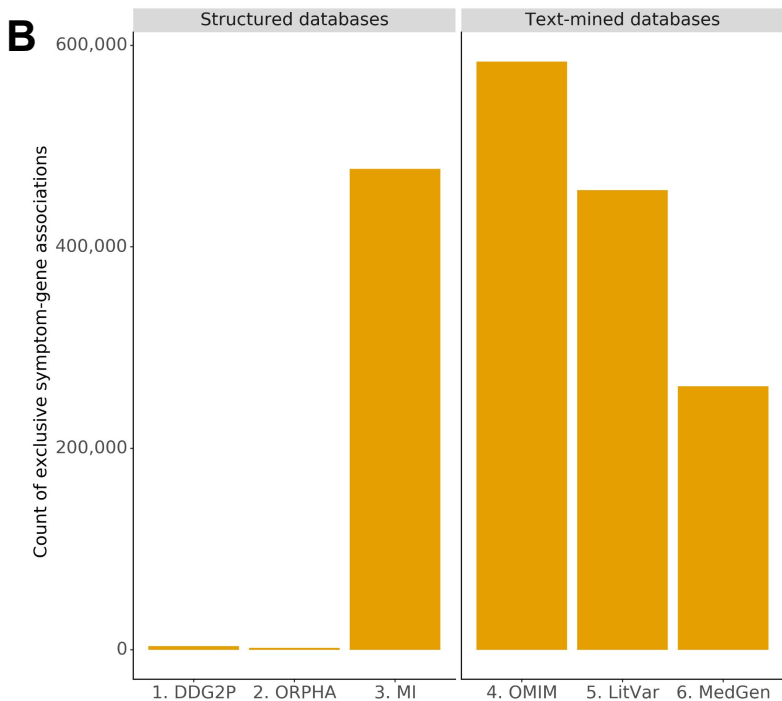
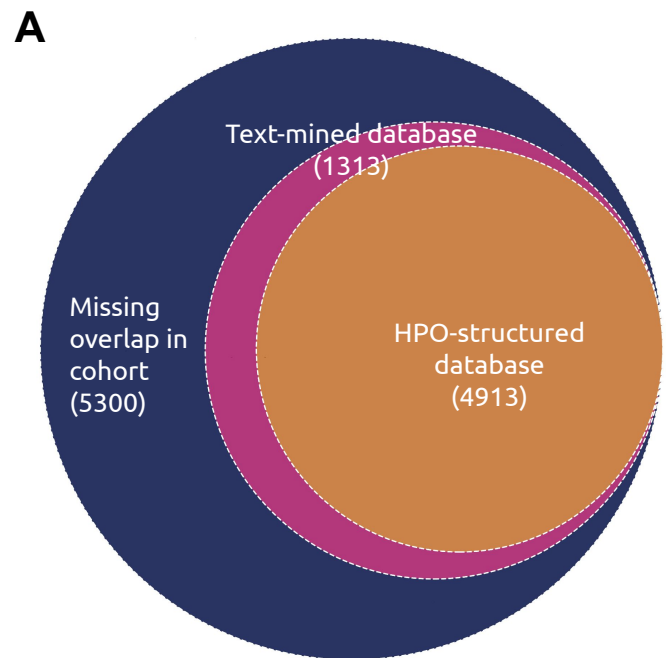
- G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
10. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–98 (2015).
  11. Louden, D. N. MedGen: NCBI’s Portal to Information on Medical Conditions with a Genetic Component. *Med. Ref. Serv. Q.* **39**, 183–191 (2020).
  12. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
  13. Röder, M., Both, A. & Hinneburg, A. Exploring the space of topic coherence measures. in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15* (ACM Press, 2015). doi:10.1145/2684822.2685324.
  14. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* vol. 3 861 Preprint at <https://doi.org/10.21105/joss.00861> (2018).
  15. Chen, Z. *et al.* PhenoApt leverages clinical expertise to prioritize candidate genes via machine learning. *Am. J. Hum. Genet.* **109**, 270–281 (2022).
  16. Zhao, M. *et al.* Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom Bioinform* **2**, lqaa032 (2020).
  17. Robinson, P. N. *et al.* Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *Am. J. Hum. Genet.* **107**, 403–417 (2020).
  18. 100,000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
  19. Larkin, J., McDermott, J., Simon, D. P. & Simon, H. A. Expert and novice performance in solving physics problems. *Science* **208**, 1335–1342 (1980).

20. Coderre, S., Mandin, H., Harasym, P. H. & Fick, G. H. Diagnostic reasoning strategies and diagnostic success. *Med. Educ.* **37**, 695–703 (2003).
21. Shen, F. *et al.* HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology. *J. Biomed. Inform.* **96**, 103246 (2019).
22. Tiddi, I. & Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **302**, 103627 (2022).
23. Zhang, Y. *et al.* High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat. Protoc.* **14**, 3426–3444 (2019).
24. Jacobsen, J. O. B. *et al.* The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol.* **40**, 817–820 (2022).
25. Yurkovich, J. T., Tian, Q., Price, N. D. & Hood, L. A systems approach to clinical oncology uses deep phenotyping to deliver personalized care. *Nat. Rev. Clin. Oncol.* **17**, 183–194 (2020).
26. Chute, C. G. Clinical Data Retrieval and Analysis. I've Seen a Case Like That Before. *Annals of the New York Academy of Sciences* vol. 670 133–140 Preprint at <https://doi.org/10.1111/j.1749-6632.1992.tb26083.x> (1992).

## Figures legends

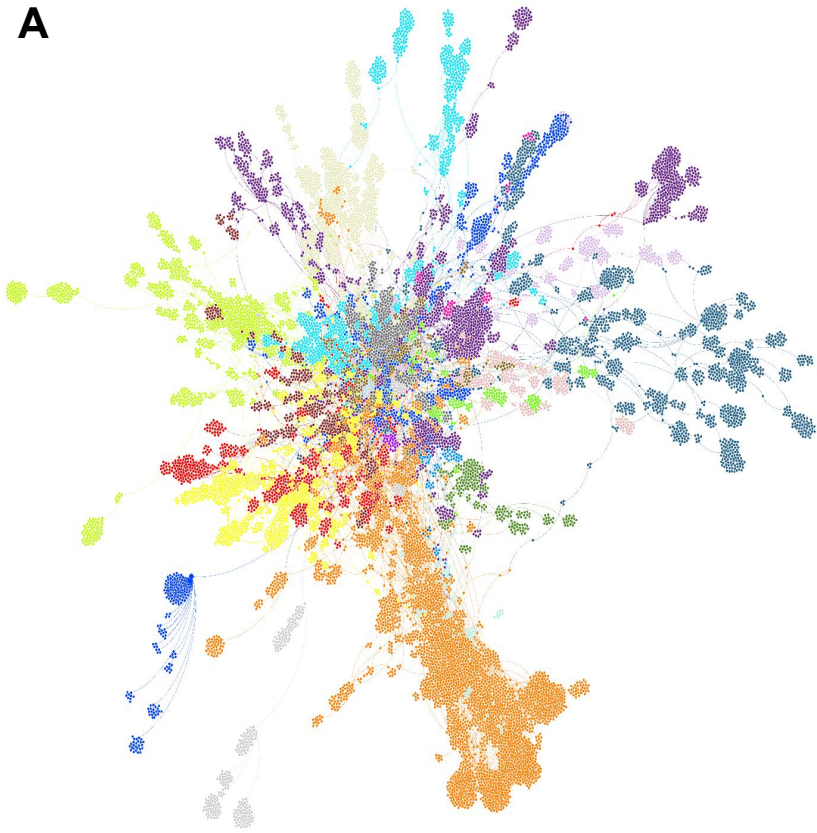
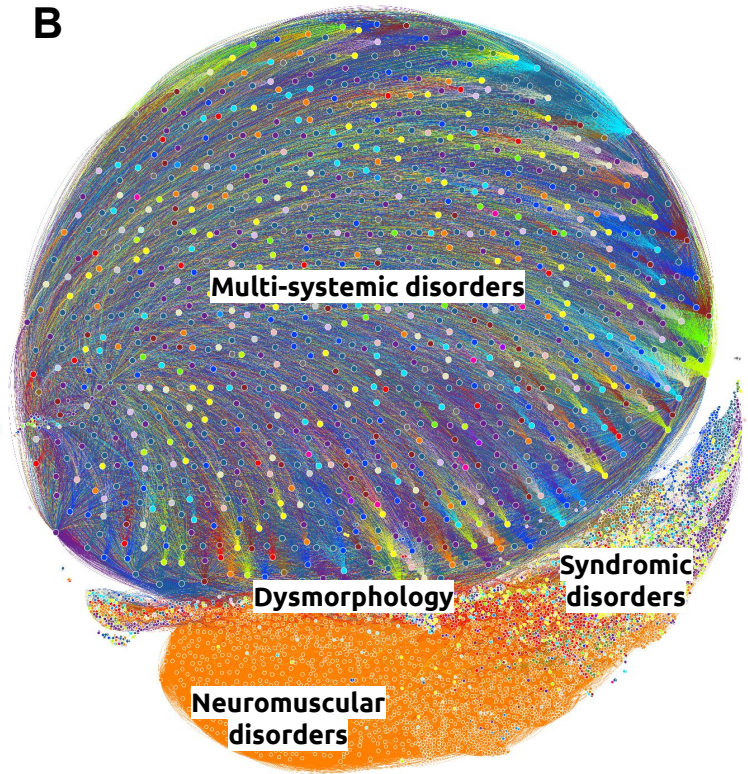
**A****B****C**

**Fig 1. The landscape of phenotyping practices from a retrospective cohort of 1,686 patients and a prospective experiment of clinical reports phenotyped by multiple physicians.** A. Treemap chart of the HPO terms frequency across the retrospective cohort. B. Violin plot of HPO term counts per clinical description for each subgroup of the cohort. C. Violin plot of HPO term counts per clinical description for each clinical report phenotyped by 12 physicians.

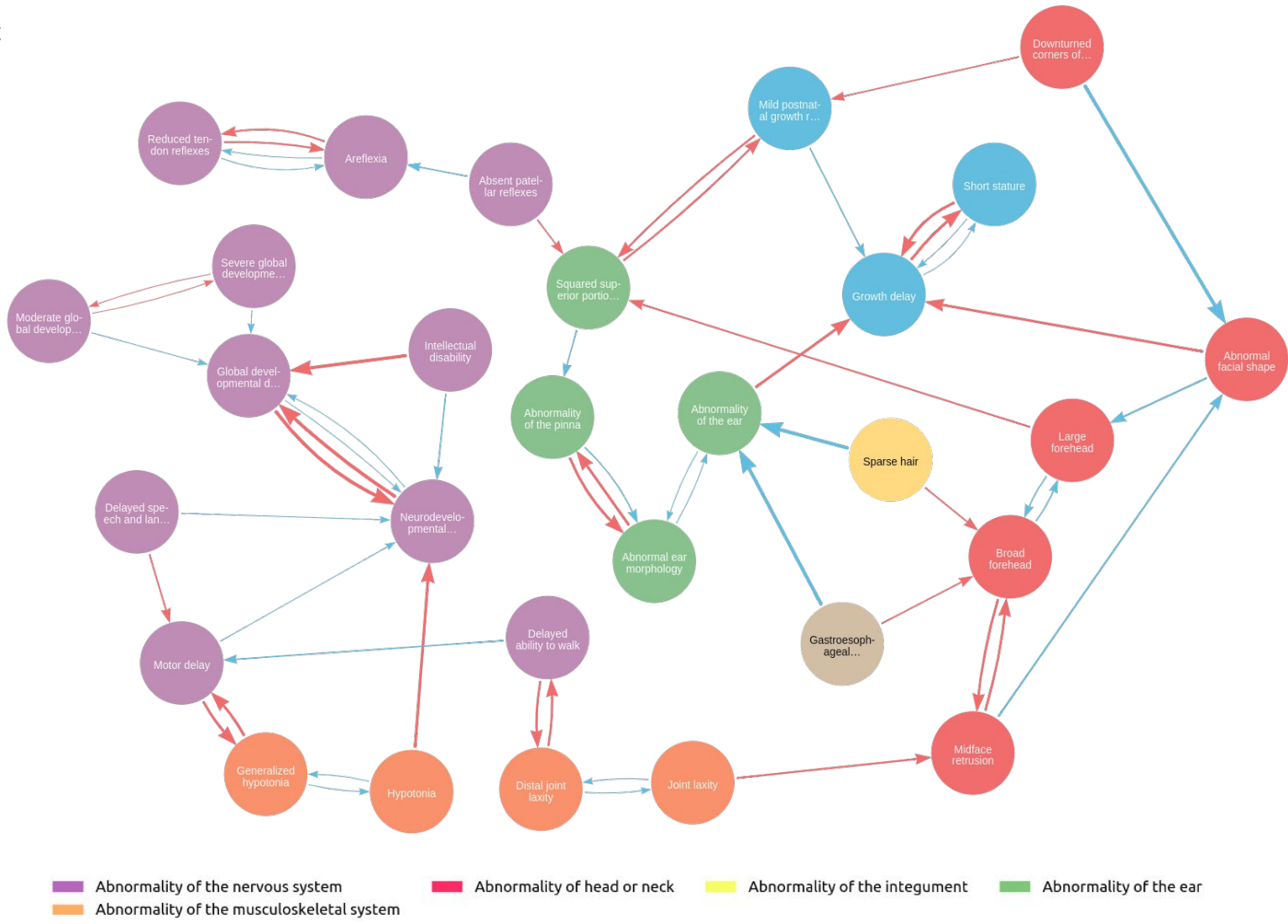


**Fig 2. Quantifying the overlap of symptoms-gene associations between the retrospective multicenter cohort of 1,686 patients and the medical literature.** A. Venn diagram of symptom-gene associations observed in cohort overlapped with public HPO-structured databases and text-mined associations in free-text databases. B. Count distribution of symptoms-gene association exclusive to each database. C. Illustration of exclusive symptom-gene associations found in Monarch Initiative database (blue) and text-mined OMIM database (green), using *KMT2D* as an example. Gray associations were unfound. D. Distribution of the mean lowest distance in the ontology between exclusive terms in the Monarch Initiative database and our text-mined OMIM database, compared to a random choice of HPO terms.



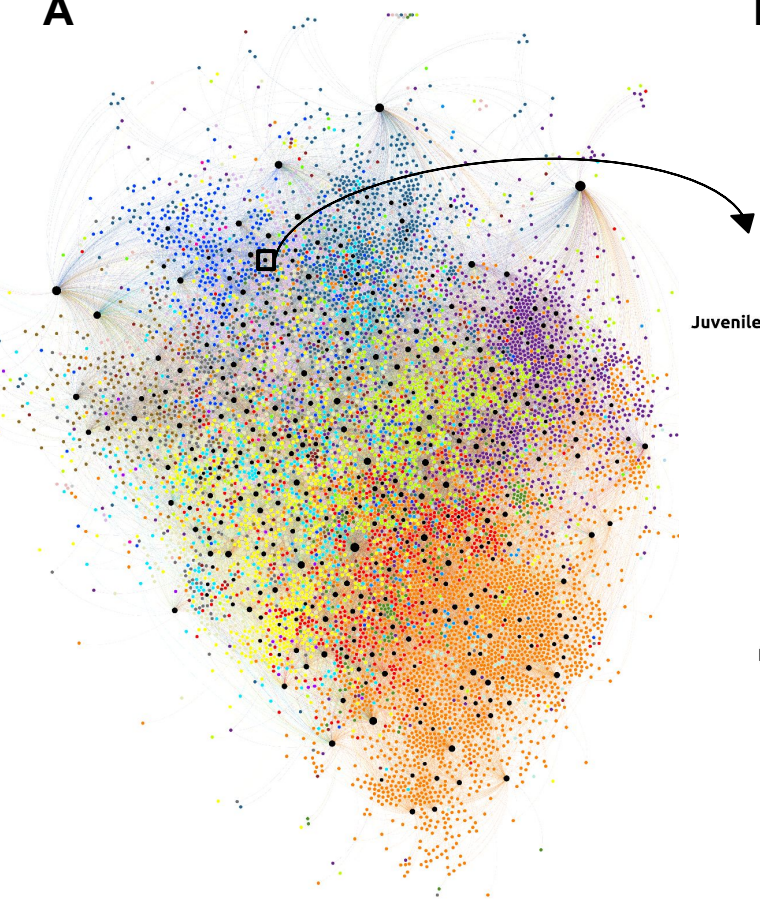
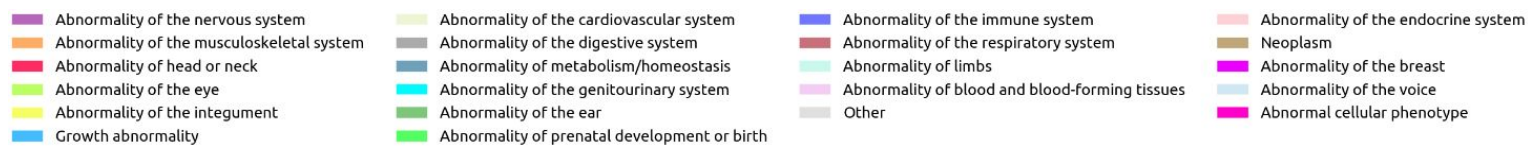
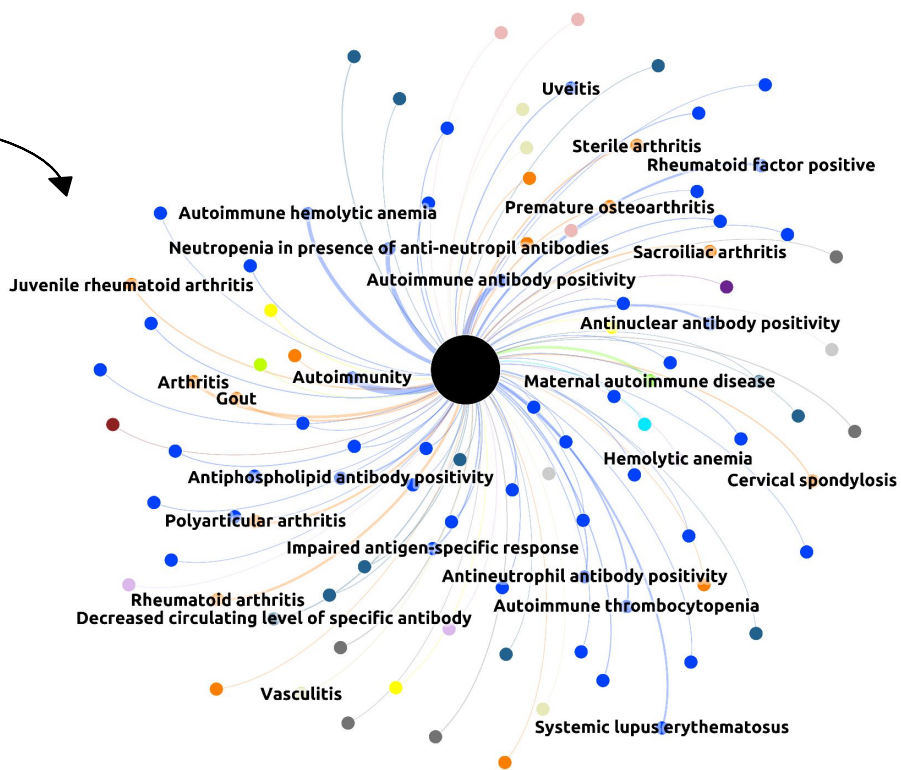
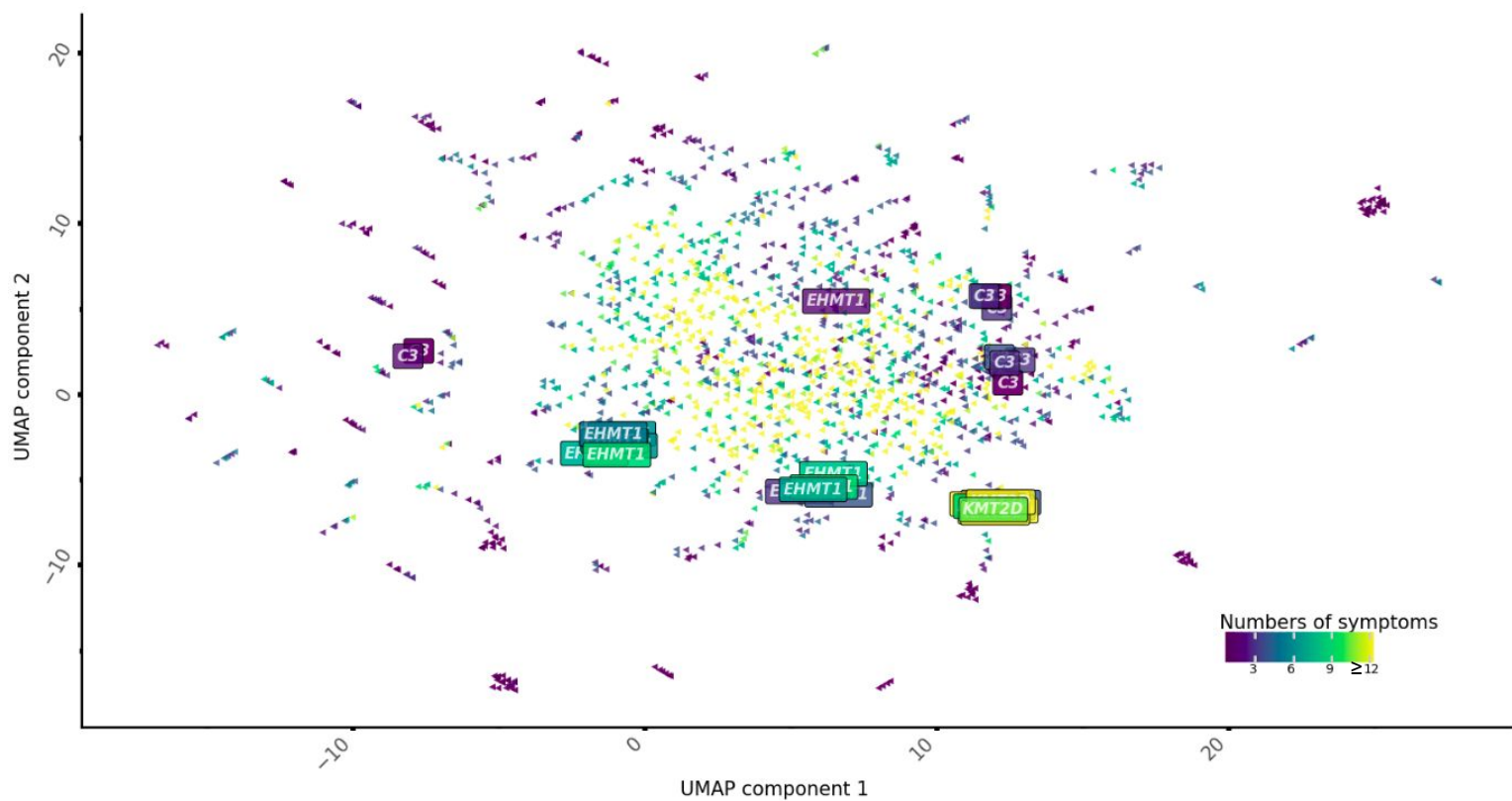
**A****B**

- |   |  |  |                                     |
|---|--|--|-------------------------------------|
| Abnormality of the nervous system         | Abnormality of the cardiovascular system     | Abnormality of the immune system               | Abnormality of the endocrine system |
| Abnormality of the musculoskeletal system | Abnormality of the digestive system          | Abnormality of the respiratory system          | Neoplasm                            |
| Abnormality of head or neck               | Abnormality of metabolism/homeostasis        | Abnormality of limbs                           | Abnormality of the breast           |
| Abnormality of the eye                    | Abnormality of the genitourinary system      | Abnormality of blood and blood-forming tissues | Abnormality of the voice            |
| Abnormality of the integument             | Abnormality of the ear                       | Other  | Abnormal cellular phenotype         |
| Growth abnormality                        | Abnormality of prenatal development or birth |  |                                     |

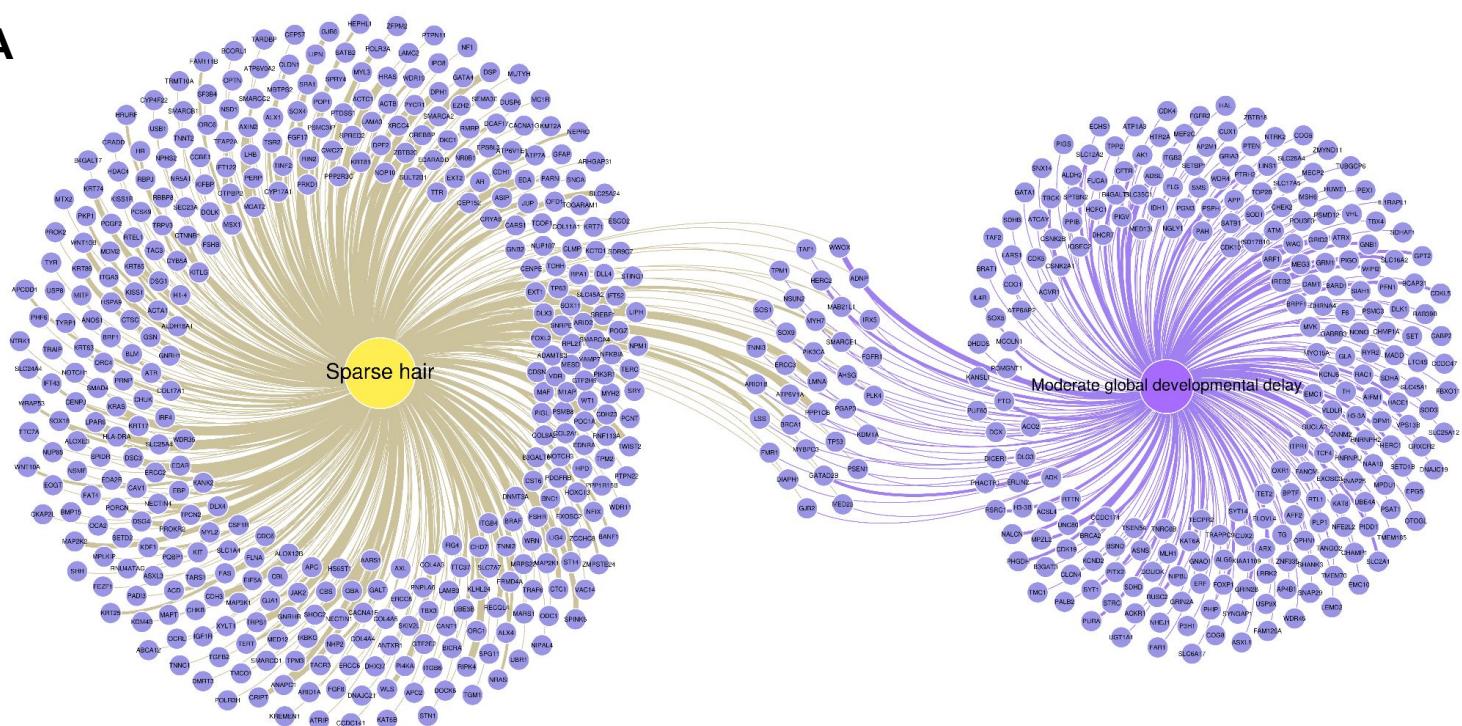
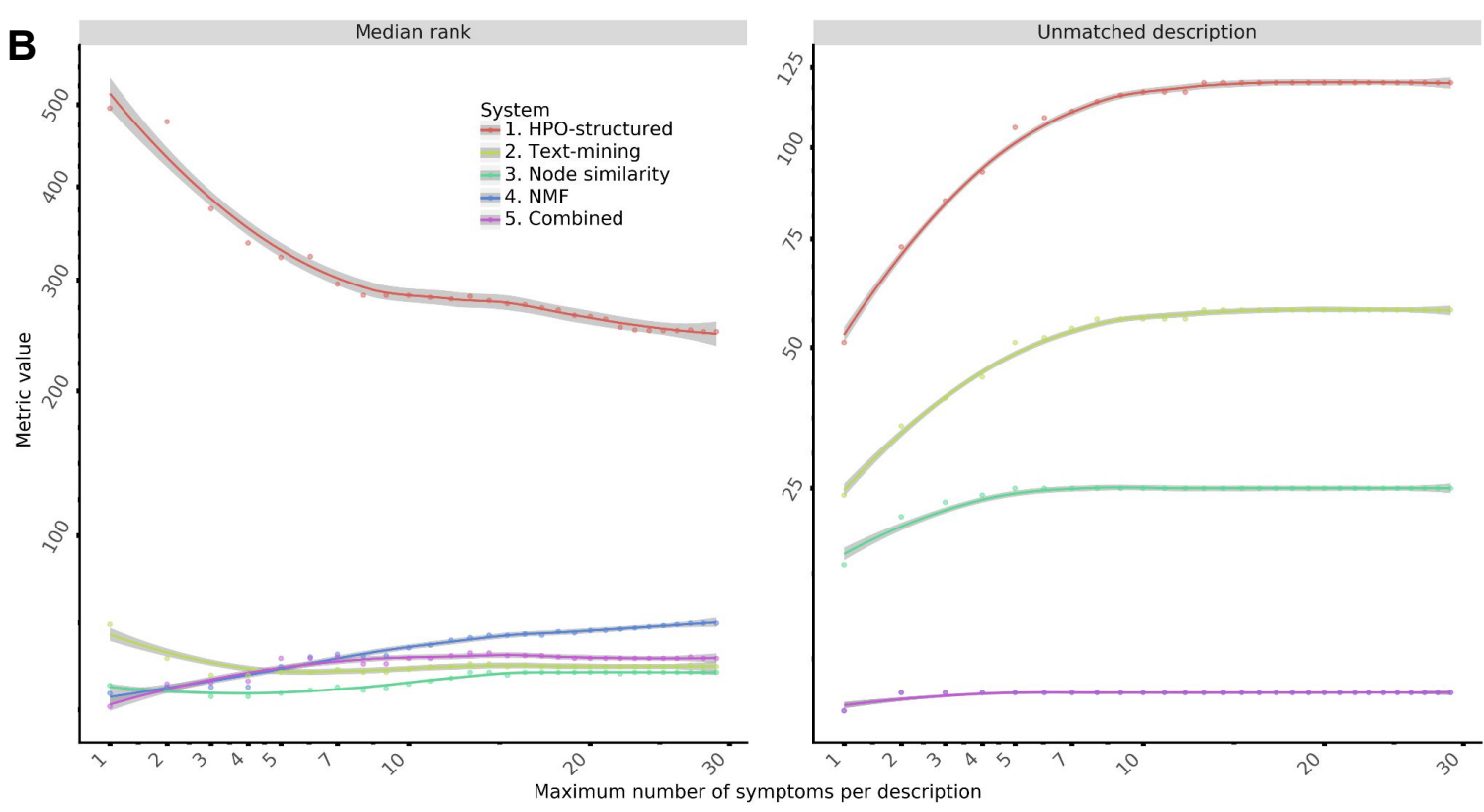
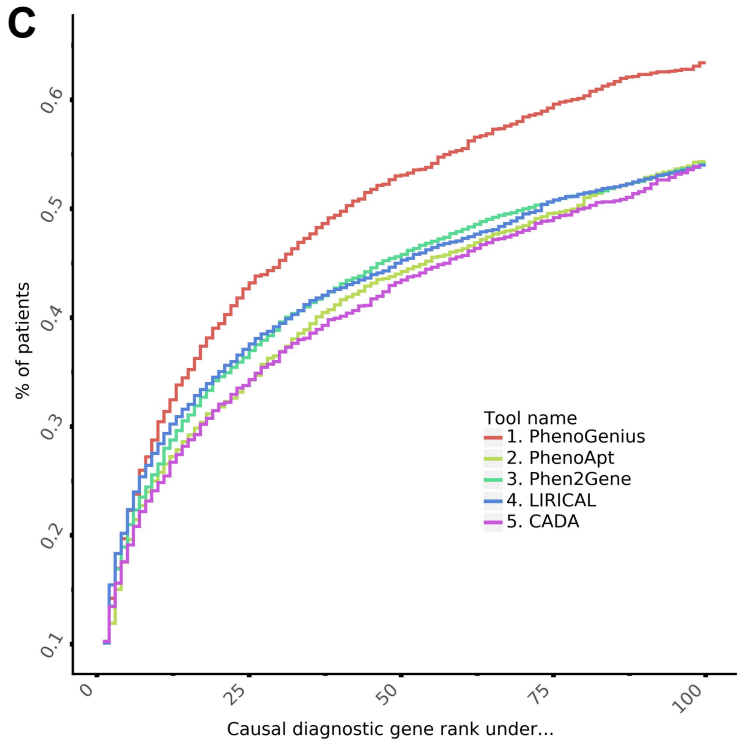
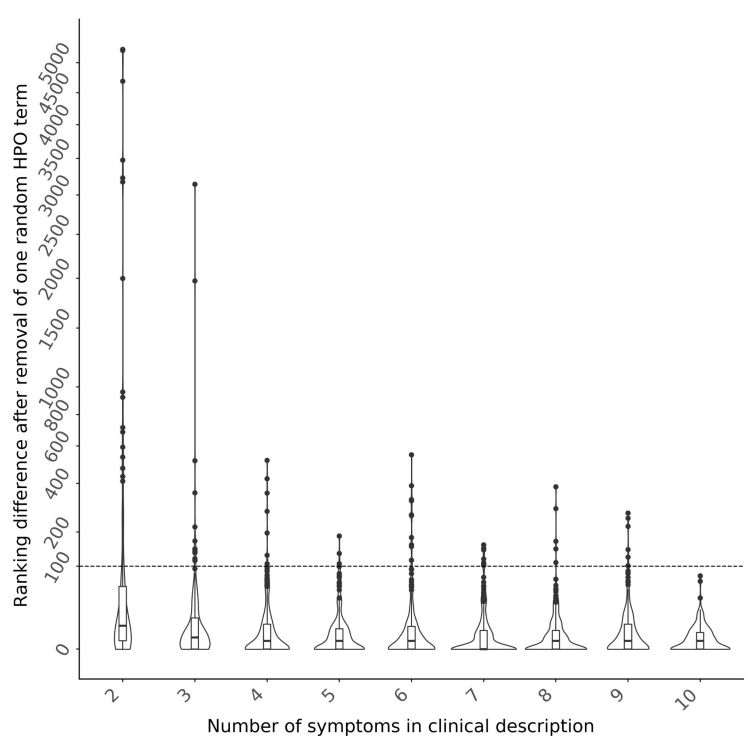
**C**



**Fig 3. Modeling symptom-symptom interaction in rare diseases using node similarity algorithms on collected symptoms-gene associations.** Node color represents the main HPO class. A. Graph visualization of symptom relationships based on the human development architecture of HPO. B. Graph visualization of symptoms relationships with node similarity > 80%. C. Illustration of symptom relationships with the Kleefstra syndrome clinical report with *EHMT1* variant, phenotyped by 12 geneticists. Blue arrows linked the closest symptom in HP ontology and red arrows the symptom with the highest node similarity among declared symptoms.

**A****B****C**

**Fig 4. Modeling symptom-symptom interactions in genetic disease using non-negative matrix factorization.** A. Visualization of symptom relationship based on 390 groups of interacting symptoms from medical literature. Group 273 is highlighted by the black box and arrow. B. Illustration of group 273 with the main symptom *Autoimmunity* (HP:0002960). For graphs in figures A and B, the line thickness is proportional to the weights of symptoms in the group. Colors correspond to the main HPO class and groups are in black. For readability, only the top 10% of symptom-group associations are displayed. C. UMAP visualization of cohort's clinical descriptions projected into the group of symptom dimension, colored by the number of symptoms. Boxes represent clinical reports description phenotyped by twelve physicians.

**A****B****C****D**



**Fig 5. Modeling symptom interactions as an efficient system for phenotype matching.** A. Illustration of the principle of phenotype matching, looking for the most connected genes to the clinical description containing two symptoms of the Kleefstra syndrome observation with the *EHMT1* variant: *Sparse hair* (HP:0008070, yellow) and *Moderate global developmental delay* (HP:0011343, purple). Line thickness is proportional to the probability score of symptom-gene associations available with joint HPO-structured and text-mined databases. B. Performance benchmark metrics of diagnostic gene prioritization ranking (median rank, left side) and phenotype matching (count of unmatched description, right side) according to a maximum number of symptoms in clinical descriptions of the cohort. C. Benchmark of a selection of state-of-the-art gene prioritization programs. The fraction of cases correctly diagnosed (y-axis) is plotted against a cumulative causal gene rank. D. Ranking differences after removing one symptom according to the number of terms in clinical descriptions.

# Materials and Methods

## Supplementary Figures and Tables for

Learning phenotypic patterns in genetic diseases by symptom interaction modeling

*Kevin Yaury*<sup>1,2,\*</sup>, *Nicolas Duforet-Frebourg*<sup>2</sup>, *Quentin Testard*<sup>1</sup>, *Sacha Beaumeunier*<sup>2</sup>, *Jerome Audoux*<sup>2</sup>,  
*Benoit Simard*<sup>3</sup>, *Dimitri Larue*<sup>2</sup>, *Michael G.B. Blum*<sup>2</sup>, *Virginie Bernard*<sup>4</sup>, *David Genevieve*<sup>5</sup>, *Denis  
Bertrand*<sup>2</sup>, *PhenoGenius consortium*, *Nicolas Philippe*<sup>2</sup>, *Julien Thevenon*<sup>1,4,\*</sup>,

1 Institute of Advanced Biosciences, Centre de recherche UGA, Inserm U 1209, CNRS UMR  
5309, Grenoble, France.

2 SeqOne Genomics, Montpellier, France.

3 OuestWare, Nantes, France.

4 Centre de référence Anomalies du développement, CHU Grenoble-Alpes, Grenoble, France.

5 Department of Medical Genetics, Rare Disease, and Personalized Medicine, IRMB, University  
of Montpellier, National Institute of Health and Medical Research, Montpellier University  
Hospital Center, Montpellier, France.

\* Corresponding authors. Email: [kevin.yaury@univ-grenoble-alpes.fr](mailto:kevin.yaury@univ-grenoble-alpes.fr) and  
[JThevenon@chu-grenoble.fr](mailto:JThevenon@chu-grenoble.fr)

### **This PDF file includes:**

Materials and Methods  
Figures S1 to S13  
Tables S1 to S4

## Materials and Methods

### Clinical data collection

We collected anonymized clinical cases from four international cohorts leading to a total of 1,686 patients with a genetic diagnosis and their clinical description in HPO terms. This cohort is composed of 307 patients gathered from the PhenoGenius consortium from Centre hospitalier universitaire (CHU) Grenoble Alpes, CHU de Dijon, CHU de Montpellier, CHU de Rennes, CHU de Brest and Hospices Civils de Lyon, 140 patients from Seo *et al.* (1), 298 patients from Trujillano *et al.* (2), and 941 from Peng *et al.* (3). We also collected clinical descriptions in HPO format from 12 clinical geneticists from 12 different French hospitals (CHU Lille, CHU Montpellier, CHU Rennes, CHU Rouen, CHU La Réunion, CH Alençon, CHU Poitiers, CHU Limoges, CH Versailles, CHU Toulouse, and CHU Tours). Each physician extracted HPO terms from the same three clinical reports of patients with different diagnostic genes (*KMT2D*, *KMD6A*, and *C3*), one case each from three physicians in French hospitals (CHU Montpellier, CHU Grenoble Alpes, and APHP). Patients or legal guardians provided informed written consent for genetic analyses in a medical setting. Consent from the clinical geneticists was obtained through a survey that also collected their responses.

### Database of medical literature

Databases were downloaded in May 2022. Human Phenotype Ontology was downloaded in OBO format from the Monarch Initiative website (<https://hpo.jax.org/>). Clinical databases in HPO format were downloaded from the Monarch Initiative website in the phenotype to genes format, EBI initiative DD2GP's (4) CSV files from <https://www.ebi.ac.uk/gene2phenotype/>, and Orphanet's XML data from <https://www.orpha.net/>. Free-text databases were downloaded through API requests for OMIM (<https://www.omim.org/>), NCBI's MedGen (5) (<https://www.ncbi.nlm.nih.gov/medgen/>) and NCBI's PubMed abstracts. For the PubMed abstracts, we used the list of all likely pathogenic and pathogenic variants from the ClinVar database (6) to select abstracts of potential interest through LitVar (7) API.

### Text matching algorithm

#### Methods

We developed a methodology to extract symptoms-gene associations based on Elasticsearch® v5.6 from free-text data in HPO terms and NCBI gene ID format. We first processed these databases to associate free-text data with the corresponding gene in JSON format. An Elasticsearch query was performed to match every HPO available in each gene-free-text related data and provide a list of HPO-gene associations per database. We limited the number of gene-HPO associations created to the top 100 ranked associated genes for an HPO.

#### Mean distance in the ontology between exclusive terms

For each exclusive symptom-gene association from the MI database, we computed ontology distance for every exclusive symptom-gene association from text-mined OMIM in a common

gene and kept the lowest distance. For each gene, we processed the mean distance in the ontology between exclusive terms of the MI database and our text-mined OMIM database. We performed the same experiment in the opposite direction, from exclusive association in text-mined OMIM database to MI database. To compare the distribution of mean distance against a random distribution, we computed ontology distance with a randomly selected HPO term instead of an exclusive term from the other database.

#### Knowledge data frame structure

A list of symptom-gene associations from each database was stored in a data frame containing 16,600 symptoms in columns and 5,235 genes in rows. Each cell includes the probability of symptom-gene association according to its overlap between databases (consensus score based on a mean, e.g. an association found in half of the databases received a score of 0.5). Databases structured in HPO format (MI, DDG2P, and Orphanet) were considered a unique resource, as DDG2P and Orphanet provided associations mostly overlapping with the MI database and with only 3,492 and 1,849 exclusive associations, respectively.

#### Node similarity

We transposed the symptoms-gene associations' data frame into a symptom-symptom association data frame based on symptoms association in the same genes. We injected all existing symptoms to symptoms relationships into a Neo4j® database v4.4.0. The similarity between all pairs of symptoms was processed using the node similarity algorithm (<https://neo4j.com/docs/graph-data-science/current/algorithms/node-similarity/>), and due to technical reasons (RAM limit due to number of combinations), for each symptom, we extracted symptoms with a similarity score  $> 0.4$  and limited to a maximum of 1500 associated symptoms. A similar pair of symptoms was reported if the similarity score was higher than 0.8.

#### Collaborative filtering

#### Methods

Using sci-kit learn v.0.24.2 Non-Negative Matrix Factorization with Nonnegative Double Singular Value Decomposition initialization, we transposed the symptoms-gene associations data frame into a symptoms-groups of symptoms association data frame based on symptoms association in the same genes. This algorithm provides the numbers of groups requested, the weights of symptom-group associations, and the weight of gene-group associations. For interpretability and illustration, we filtered symptoms-group association keeping only the 10% highest l2-normalized weight of symptoms-group association ( $>0.04$ ). In phenotype matching evaluation, we use the complete symptoms-group associations.

#### Identification of an optimal number of symptoms group

We applied a coherence score metric to processed groups of symptoms to select their optimal number. Using gensim v4.2 implementations of the coherence topic evaluation (<https://radimrehurek.com/gensim/models/coherencemodel.html>), we sought the range of group numbers with the highest coherence score and also the lowest coefficient of variation using five random state initialization. We looked for consistency of coherence among different random



states to reproduce the same performance with additional data or updates. This is necessary for clinical implementation.

### UMAP and clustering

Using sci-kit learn's v0.24.2 agglomerative clustering implementation (affinity="euclidean" and linkage="ward" parameters), we obtained hierarchical clusters. Data were normalized per clinical observations and also per group of symptoms. 75 clusters of clinical descriptions were retained to avoid single observation clusters. Dendrograms were obtained using Scipy v1.8. hierarchy module. UMAP visualization was performed using the umap v0.5.3 module, with the following parameters: neighbours = 3 & minimum distance = 0.9 for 390 groups of symptom dimension and neighbours = 2 & minimum distance = 0.9 for 16,600 symptoms dimension. We determined these parameters after observing dispersion and clustering of clinical descriptions according to a range of neighbors from (2 to 5) and minimum distance (0.1 to 0.99).

### Graph visualization

Exploration plots were processed using python's plotnine v0.9 package. Graph visualization was obtained using software Gephi v0.9 (<https://gephi.org/>) with ForceAtlas2 (8) and Neo4j Bloom v2.3. Retina (<https://ouestware.gitlab.io/retina/beta/>) was used to provide users with a visual browser of graphs.

### Phenotype matching

A phenotype match occurred if at least one symptom in the clinical description was related to the diagnostic gene in the database. We developed a phenotype matching system that matches clinical descriptions with lists of symptoms-gene associations available in the knowledge data frame structure. According to the combination of symptoms from clinical descriptions, the data frame was filtered to contain only selected symptoms or groups of symptoms columns. The sum of the consensus score per row or gene was processed, and genes ranked according to the sum score in descending order. In the case of equal scoring, we applied the worst rank to all equal genes. The evaluation of this phenotype matching using databases in HPO format and text-mined associations used only symptoms declared in the clinical description. The phenotype matching system based on symptoms similarity used declared symptoms and added a virtual symptom containing all highly similar symptoms sharing its weight. The method based on NMF's collaborative filtering projected symptoms into 390 groups of interacting symptoms dimension using a trained model. Each gene is also projected in the 390 groups dimension with different weights. The ranking is calculated based on the normalized Euclidean distance between the 390 group projection of each gene and the patient phenotypes. The gene with the highest distance gets the best ranking.

### Comparisons of phenotype-driven gene prioritization systems

We benchmarked the performance of four different phenotype-driven gene prioritization algorithms: PhenoApt, Phen2Gene, CADA, and LIRICAL (3, 9–11). As each tool provides a different maximum limit for the gene ranking list, we performed our diagnostic performance evaluation based on the top 100 cumulative causal gene ranks to be interpretable. PhenoApt (<https://www.phenoapt.org/API>) and Phen2Gene (<https://phen2gene.wglab.org/api>) evaluations were processed using the software's API in May 2022. CADA

(<https://github.com/Chengyao-Peng/CADA>, unique release) and LIRICAL (<https://github.com/TheJacksonLaboratory/LIRICAL>, v.1.3.4) evaluations were processed using the desktop version via GitHub.

## Statistics

Statistical metrics (Kolmogorov-Smirnov test, Spearman correlation coefficient, Fisher exact test) were obtained using Python's SciPy module v1.8. Benjamini Hochberg correction was obtained using the multiply package v0.16.

## Methods references

1. G. H. Seo, T. Kim, I. H. Choi, J.-Y. Park, J. Lee, S. Kim, D.-G. Won, A. Oh, Y. Lee, J. Choi, H. Lee, H. G. Kang, H. Y. Cho, M. H. Cho, Y. J. Kim, Y. H. Yoon, B.-L. Eun, R. J. Desnick, C. Keum, B. H. Lee, Diagnostic yield and clinical utility of whole exome sequencing using an automated variant prioritization system, *EVIDENCE. Clin. Genet.* **98**, 562–570 (2020).
2. D. Trujillano, A. M. Bertoli-Avella, K. Kumar Kandaswamy, M. E. Weiss, J. Köster, A. Marais, O. Paknia, R. Schröder, J. M. Garcia-Aznar, M. Werber, O. Brandau, M. Calvo Del Castillo, C. Baldi, K. Wessel, S. Kishore, N. Nahavandi, W. Eyaid, M. T. Al Rifai, A. Al-Rumayyan, W. Al-Twaijri, A. Alothaim, A. Alhashem, N. Al-Sannaa, M. Al-Balwi, M. Alfadhel, A. Rolfs, R. Abou Jamra, Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* **25**, 176–182 (2017).
3. C. Peng, S. Dieck, A. Schmid, A. Ahmad, A. Knaus, M. Wenzel, L. Mehnert, B. Zirn, T. Haack, S. Ossowski, M. Wagner, T. Brunet, N. Ehmke, M. Danyel, S. Rosnev, T. Kamphans, G. Nadav, N. Fleischer, H. Fröhlich, P. Krawitz, CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genom Bioinform.* **3**, lqab078 (2021).
4. A. Thormann, M. Halachev, W. McLaren, D. J. Moore, V. Svinti, A. Campbell, S. M. Kerr, M. Tischkowitz, S. E. Hunt, M. G. Dunlop, M. E. Hurles, C. F. Wright, H. V. Firth, F. Cunningham, D. R. FitzPatrick, Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
5. D. N. Loudon, MedGen: NCBI's Portal to Information on Medical Conditions with a Genetic Component. *Med. Ref. Serv. Q.* **39**, 183–191 (2020).
6. M. J. Landrum, S. Chitipiralla, G. R. Brown, C. Chen, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Kaur, C. Liu, V. Lyoshin, Z. Maddipatla, R. Maiti, J. Mitchell, N. O'Leary, G. R. Riley, W. Shi, G. Zhou, V. Schneider, D. Maglott, J. B. Holmes, B. L. Kattman, ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
7. A. Allot, Y. Peng, C.-H. Wei, K. Lee, L. Phan, Z. Lu, LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Res.* **46**, W530–W536 (2018).
8. M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One.* **9**, e98679 (2014).
9. Z. Chen, Y. Zheng, Y. Yang, Y. Huang, S. Zhao, H. Zhao, C. Yu, X. Dong, Y. Zhang, L. Wang, Z.

- Zhao, S. Wang, Y. Yang, Y. Ming, J. Su, G. Qiu, Z. Wu, T. J. Zhang, N. Wu, PhenoApt leverages clinical expertise to prioritize candidate genes via machine learning. *Am. J. Hum. Genet.* **109**, 270–281 (2022).
10. M. Zhao, J. M. Havrilla, L. Fang, Y. Chen, J. Peng, C. Liu, C. Wu, M. Sarmady, P. Botas, J. Isla, G. J. Lyon, C. Weng, K. Wang, Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom Bioinform.* **2**, lqaa032 (2020).
  11. P. N. Robinson, V. Ravanmehr, J. O. B. Jacobsen, D. Danis, X. A. Zhang, L. C. Carmody, M. A. Gargano, C. L. Thaxton, UNC Biocuration Core, G. Karlebach, J. Reese, M. Holtgrewe, S. Köhler, J. A. McMurry, M. A. Haendel, D. Smedley, Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *Am. J. Hum. Genet.* **107**, 403–417 (2020).

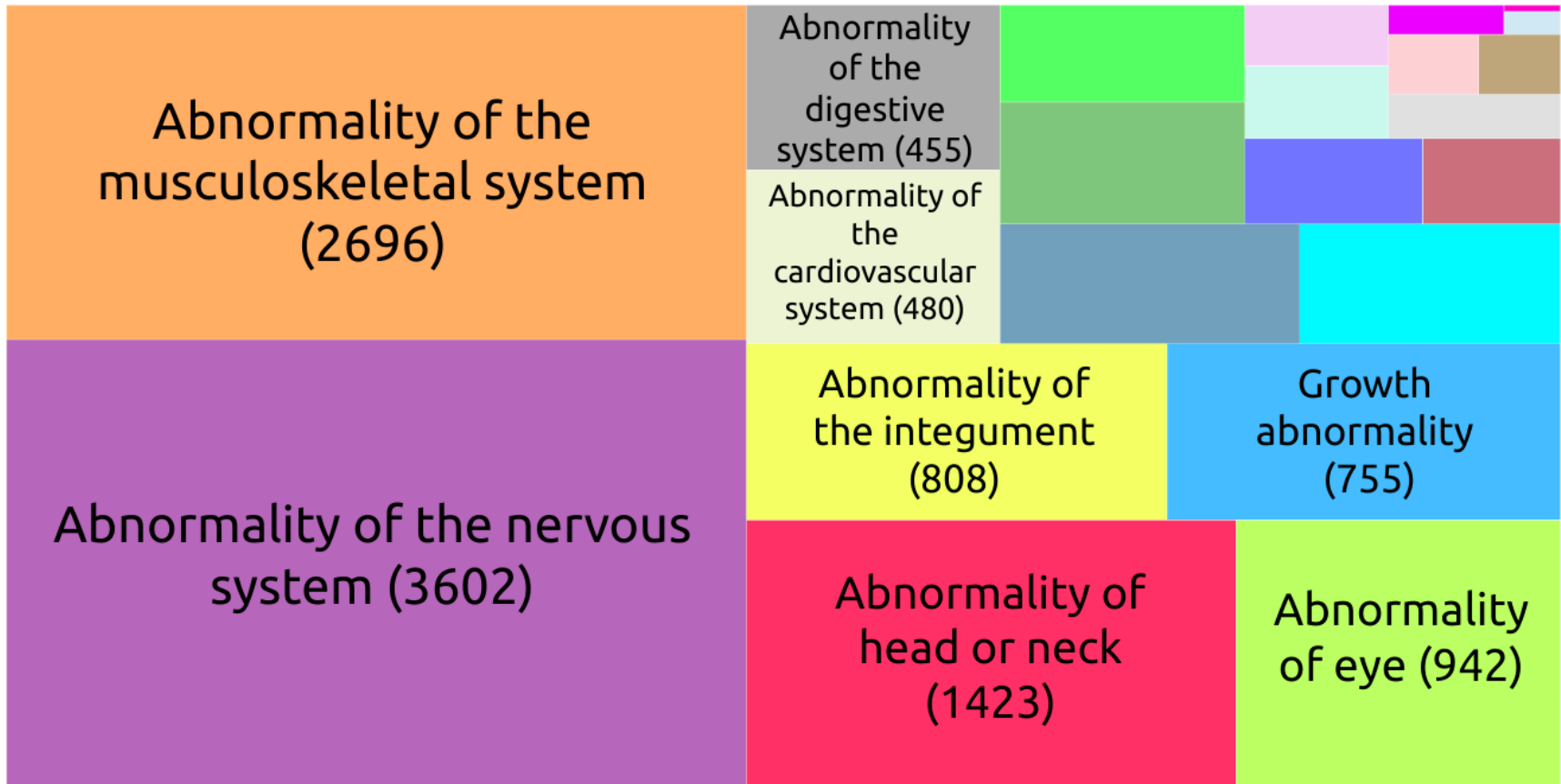
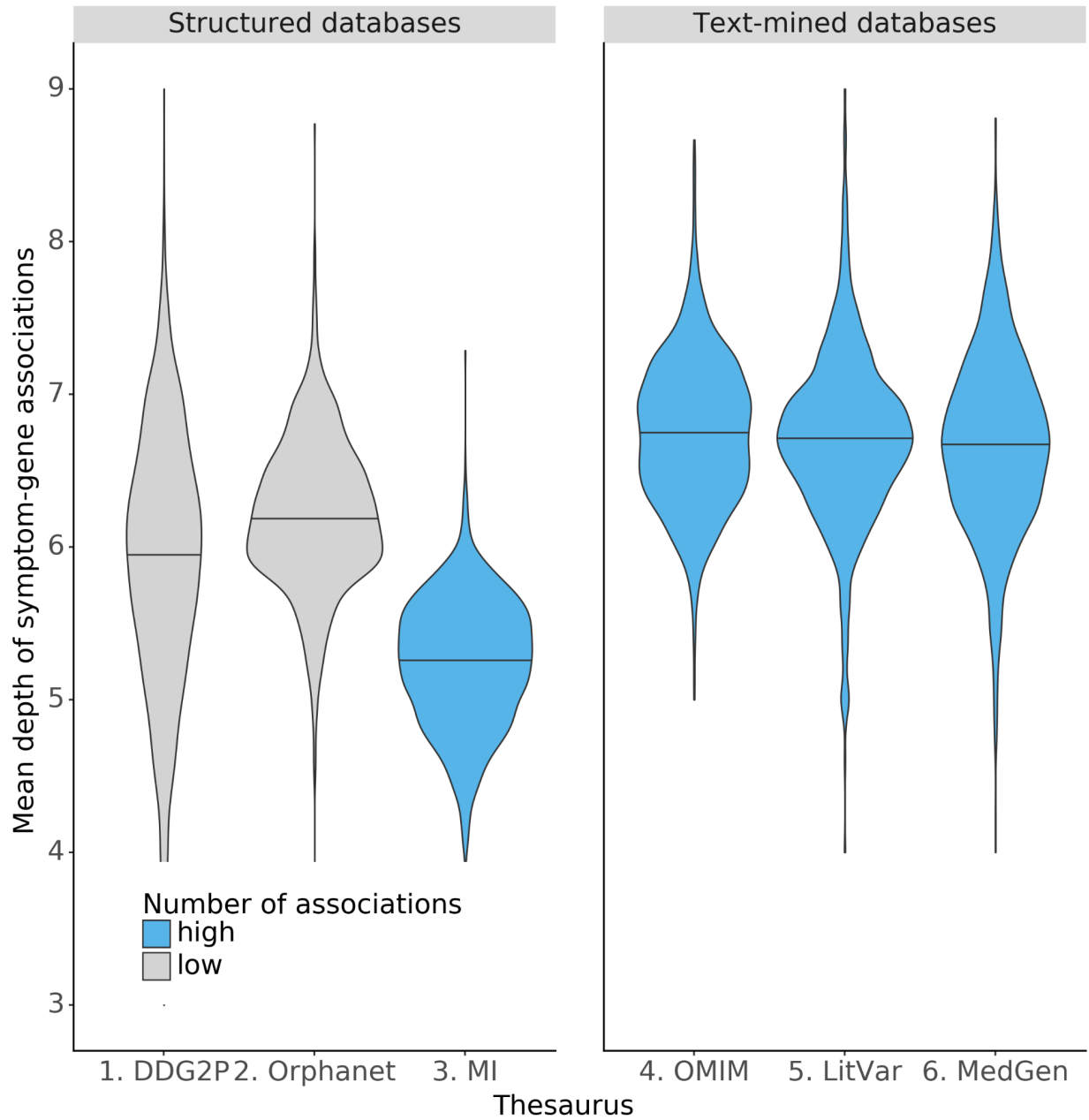
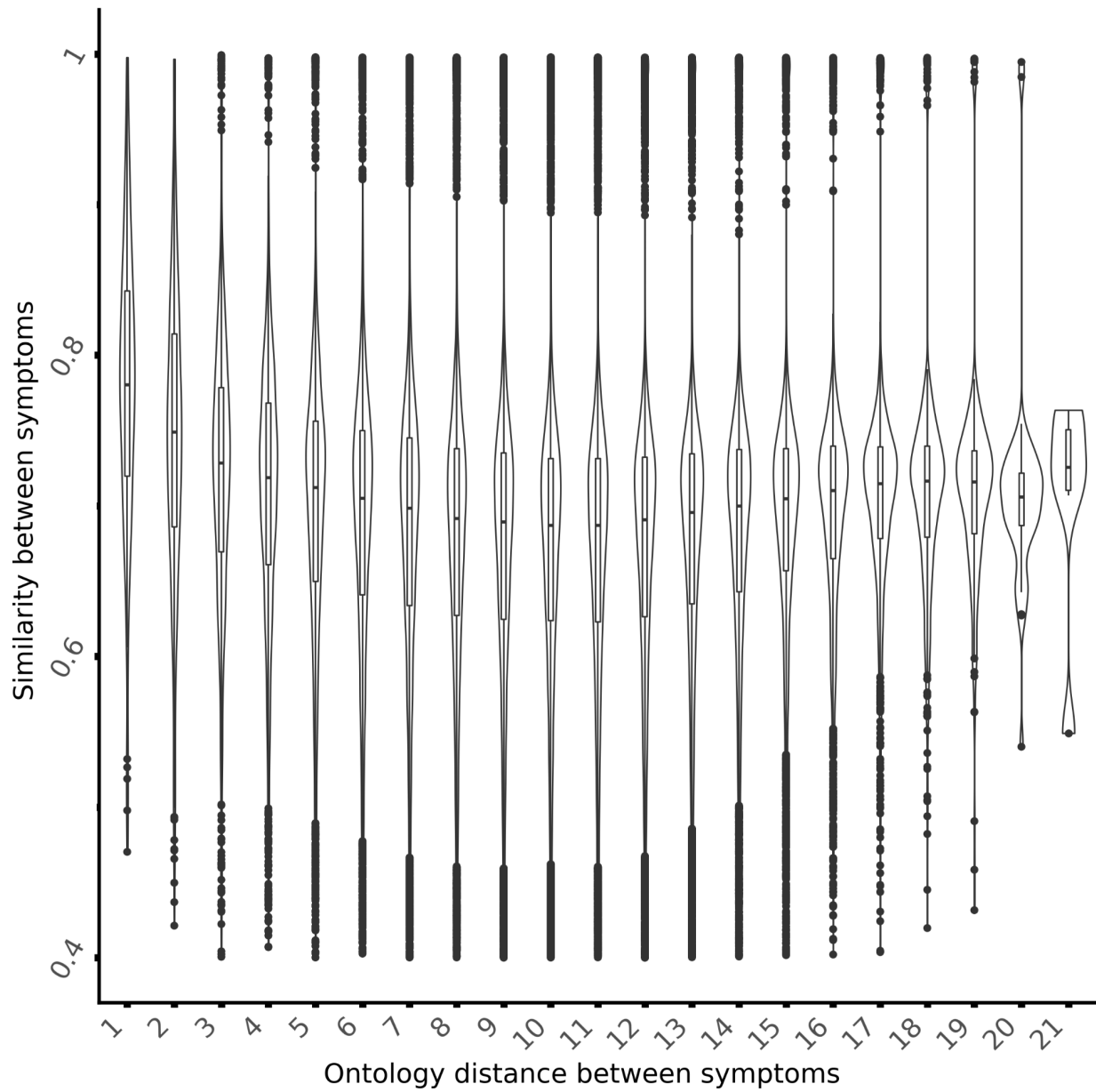


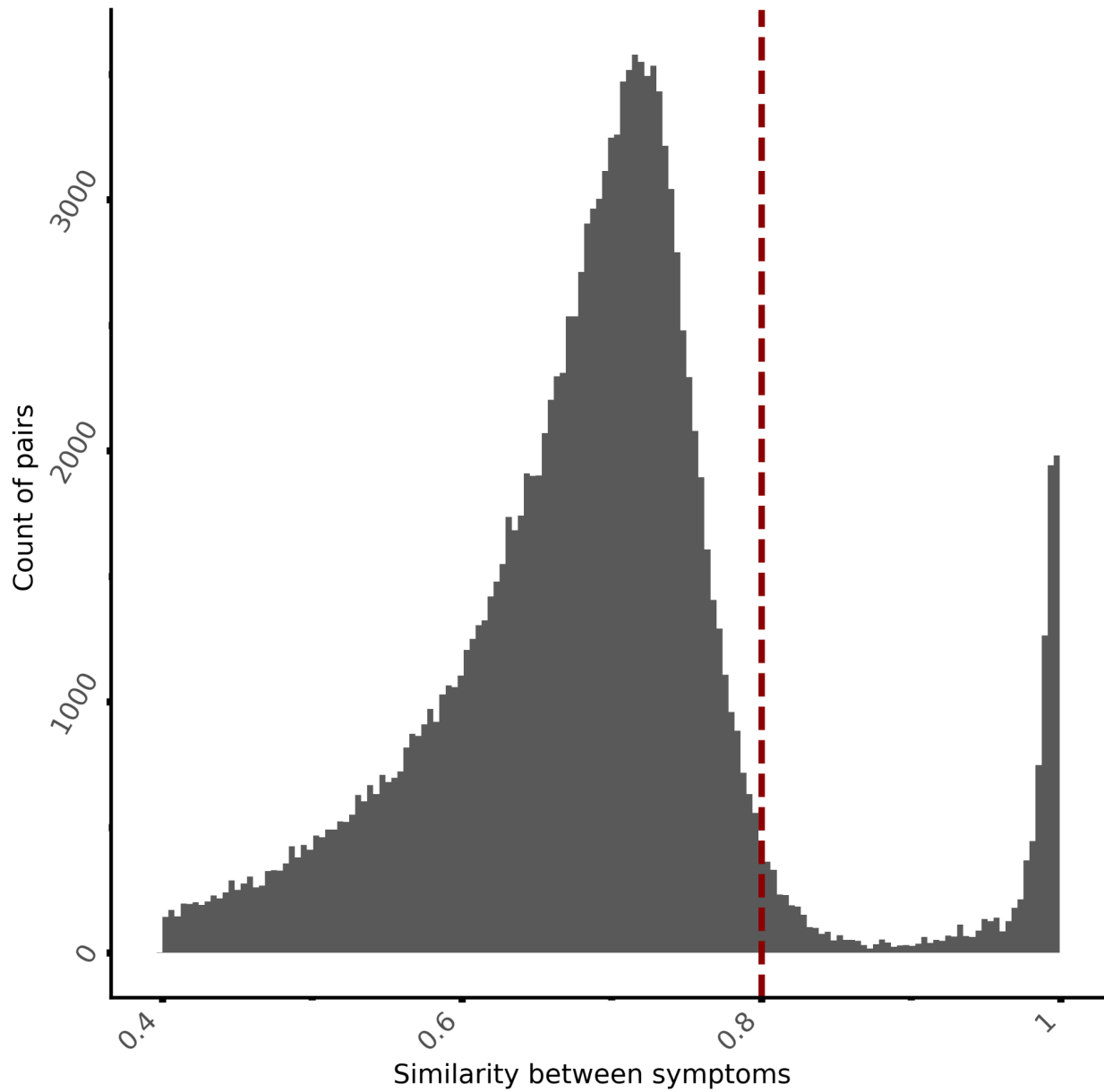
Fig. S1. A. Treemap chart of the HPO terms in the cohort per main class ontology.



**Fig. S2. Violin plot of the mean depth of HPO terms from root ontology according to each database.**

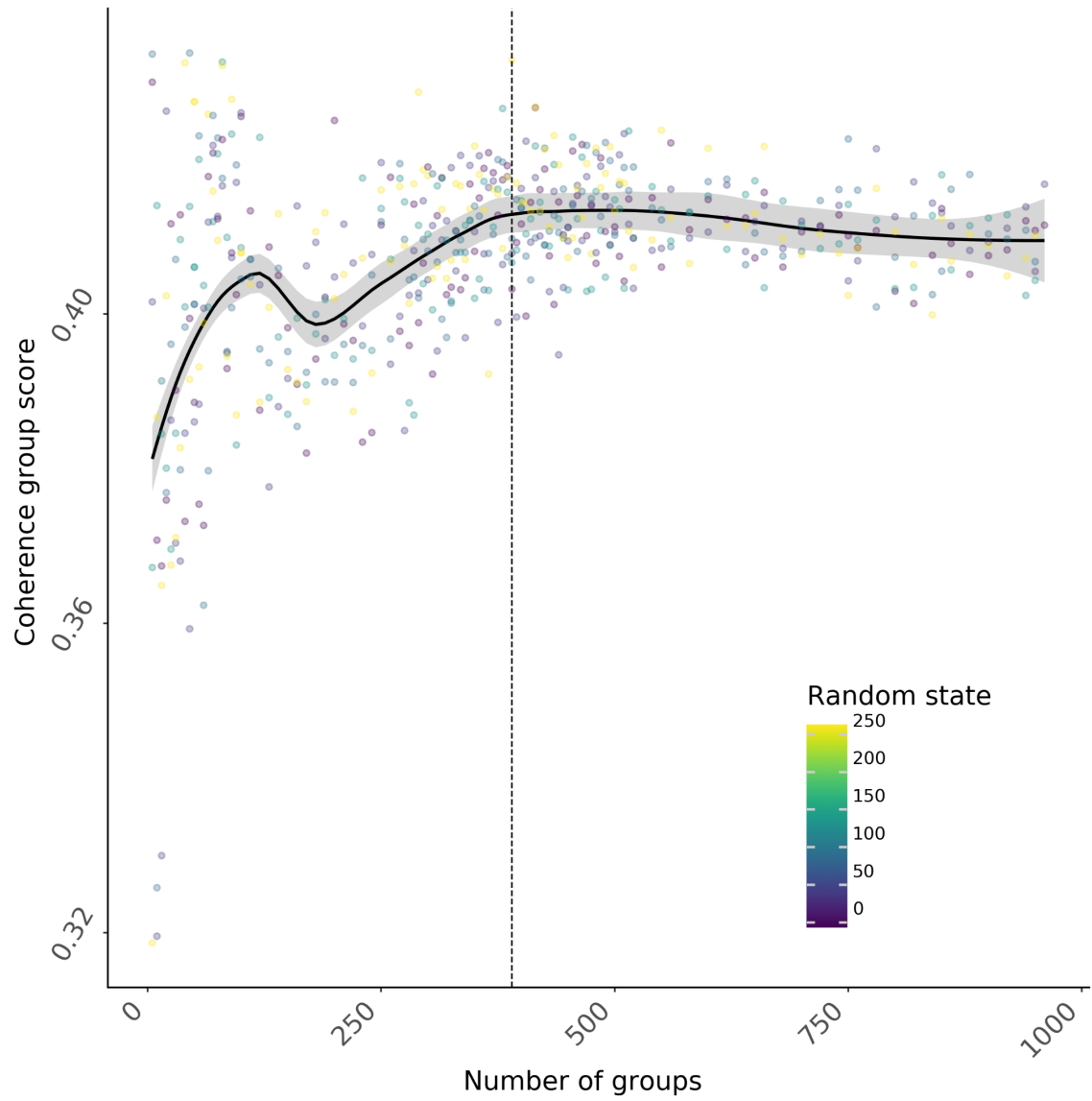


**Fig. S3. Violin plot of pair of symptoms similarity score according to the ontology distance.**



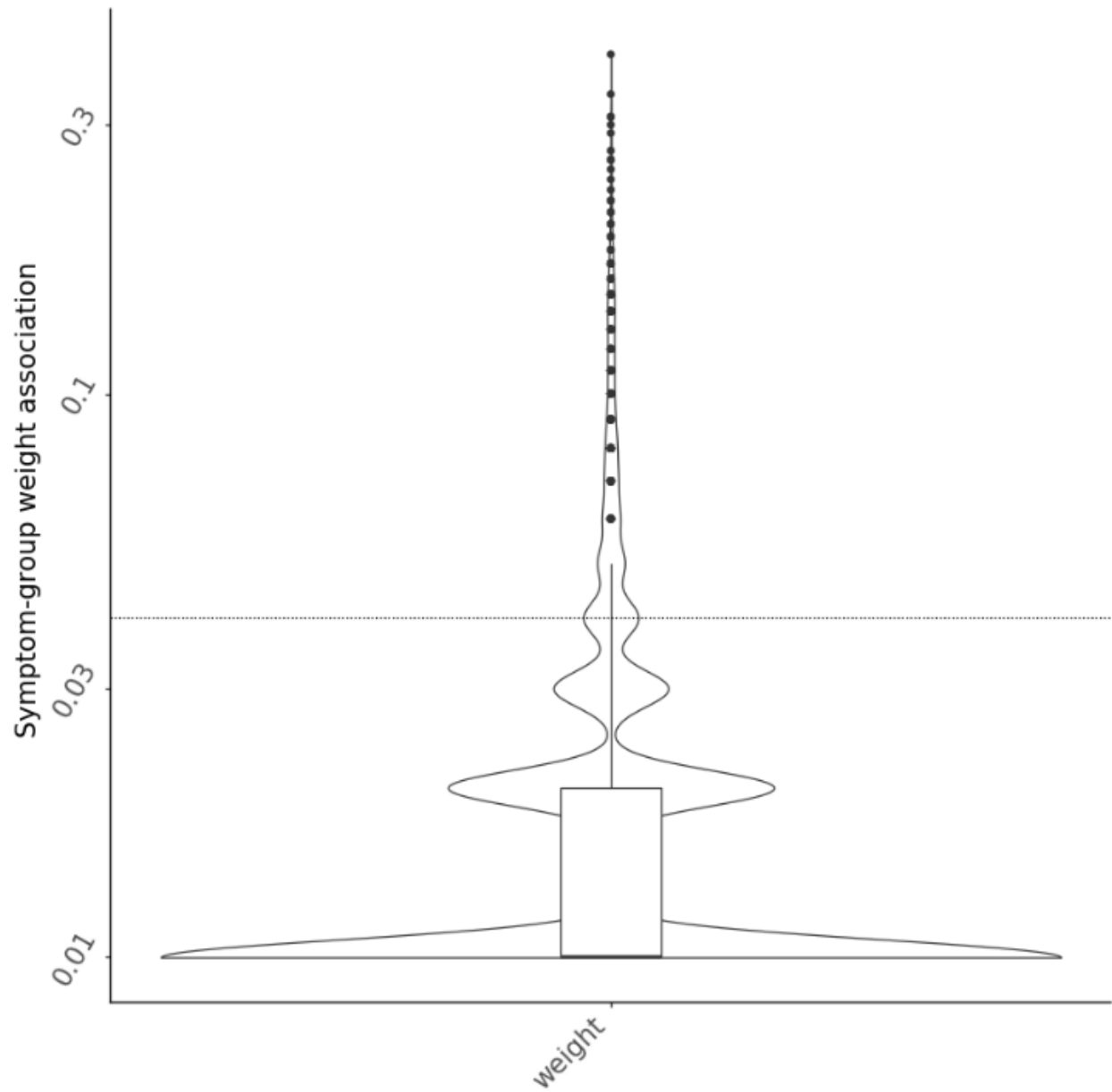
**Fig. S4. Distribution of 1% subsampling of similarity pair score.**

The dashed red line corresponds to the 80% similarity threshold and represents 10% of similarity pairs.



**Fig. S5. Topic coherence evaluation to determine the optimal number of groups.**  
Colors represent iterative training experiments according to random state initialization.

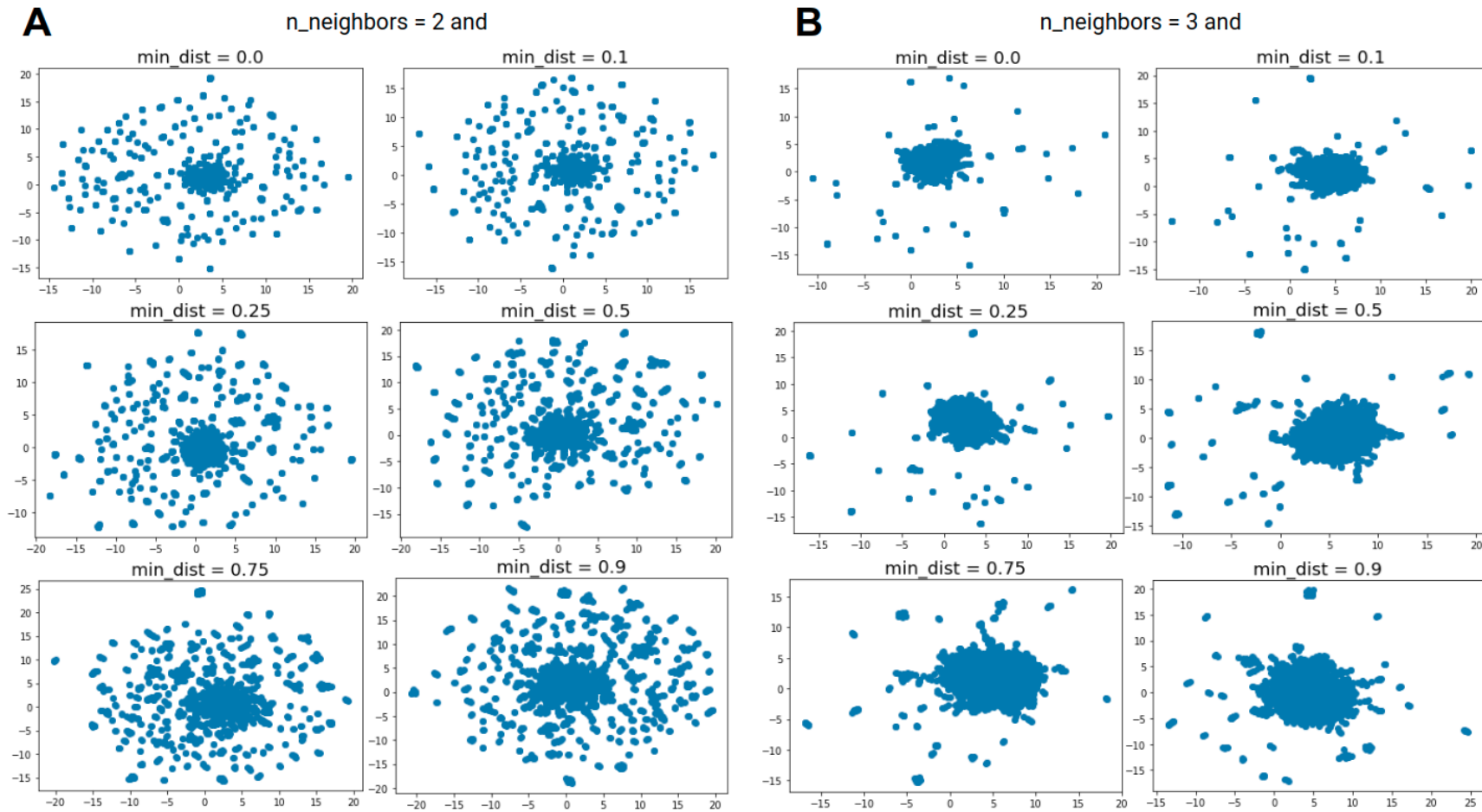




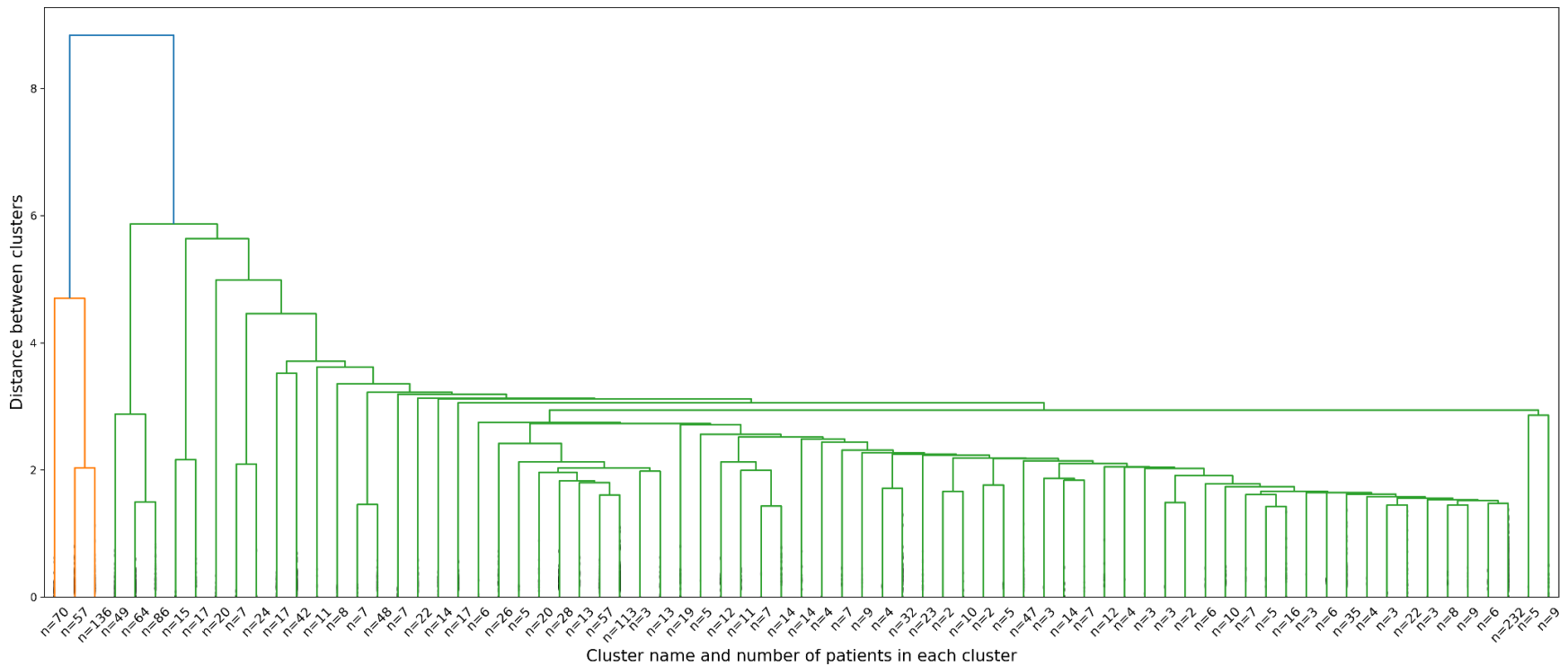
**Fig. S6. Violin plot of symptom-group weight association.**

Dashed horizontal line corresponds to the top 10% of symptom-group associations normalized weight threshold ( $>0.04$ ).

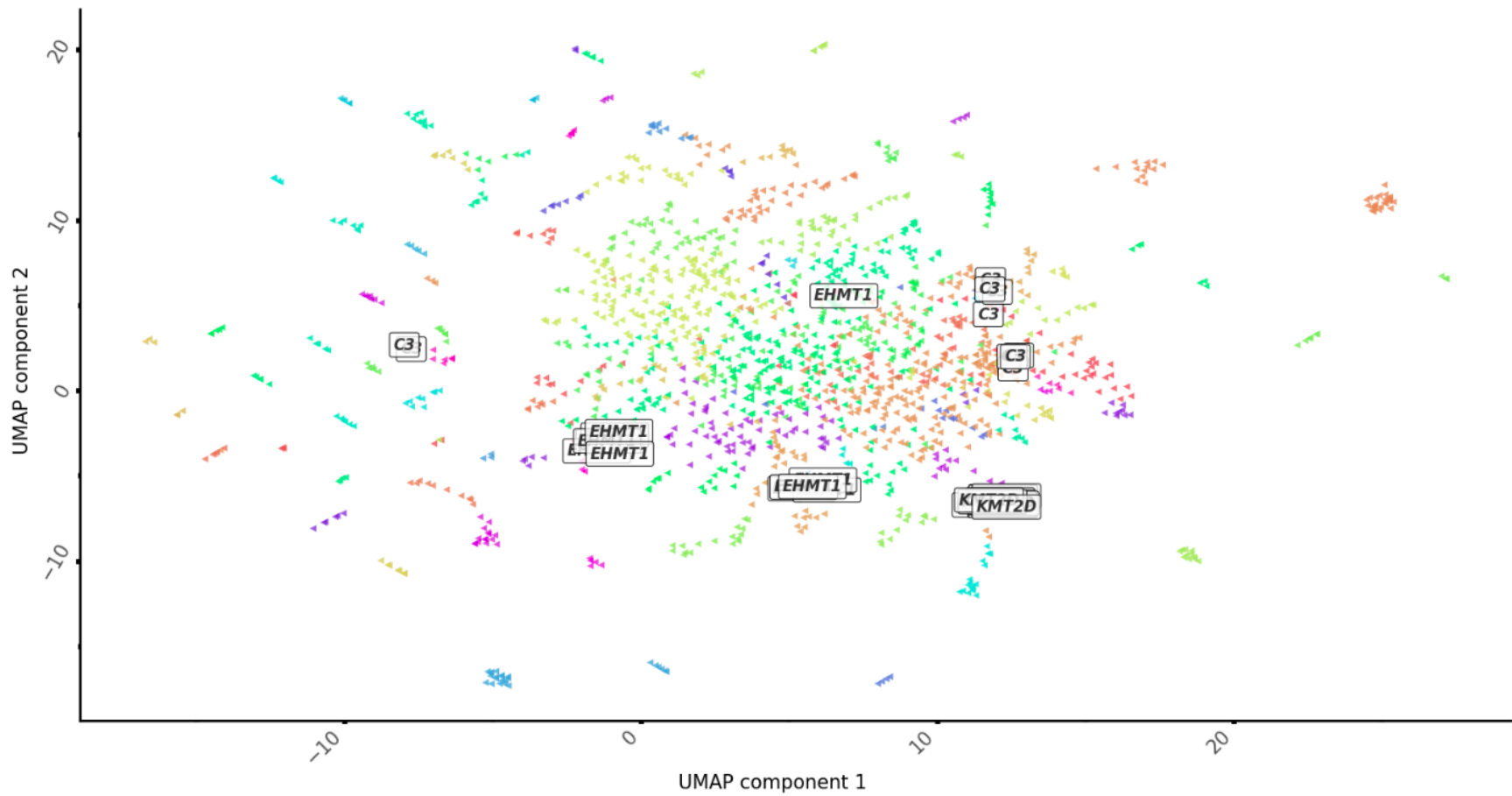




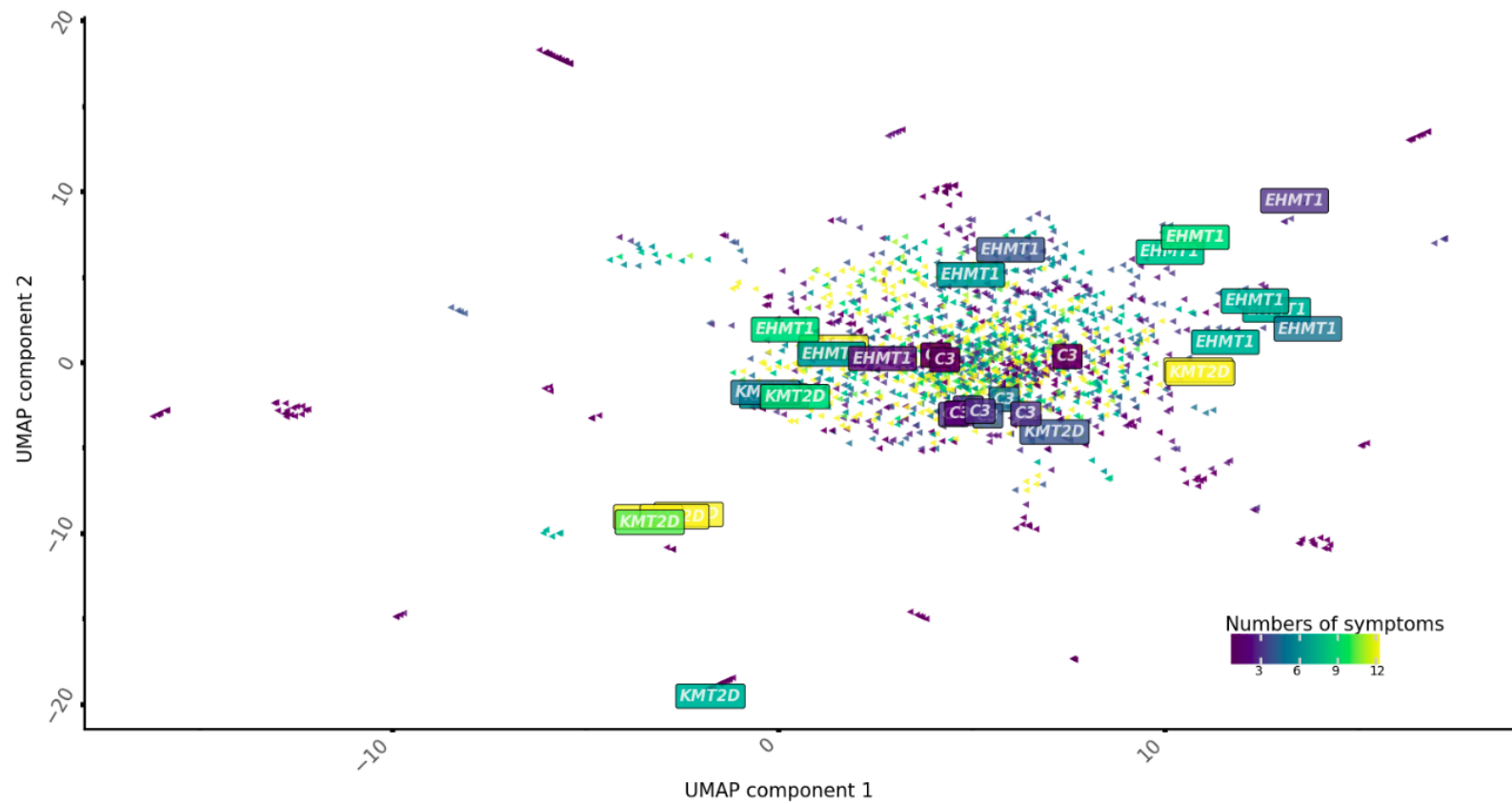
**Fig. S8. UMAP visualization of clinical description from the retrospective cohort of 1,686 cases using the 16,600 symptoms from HPO.**



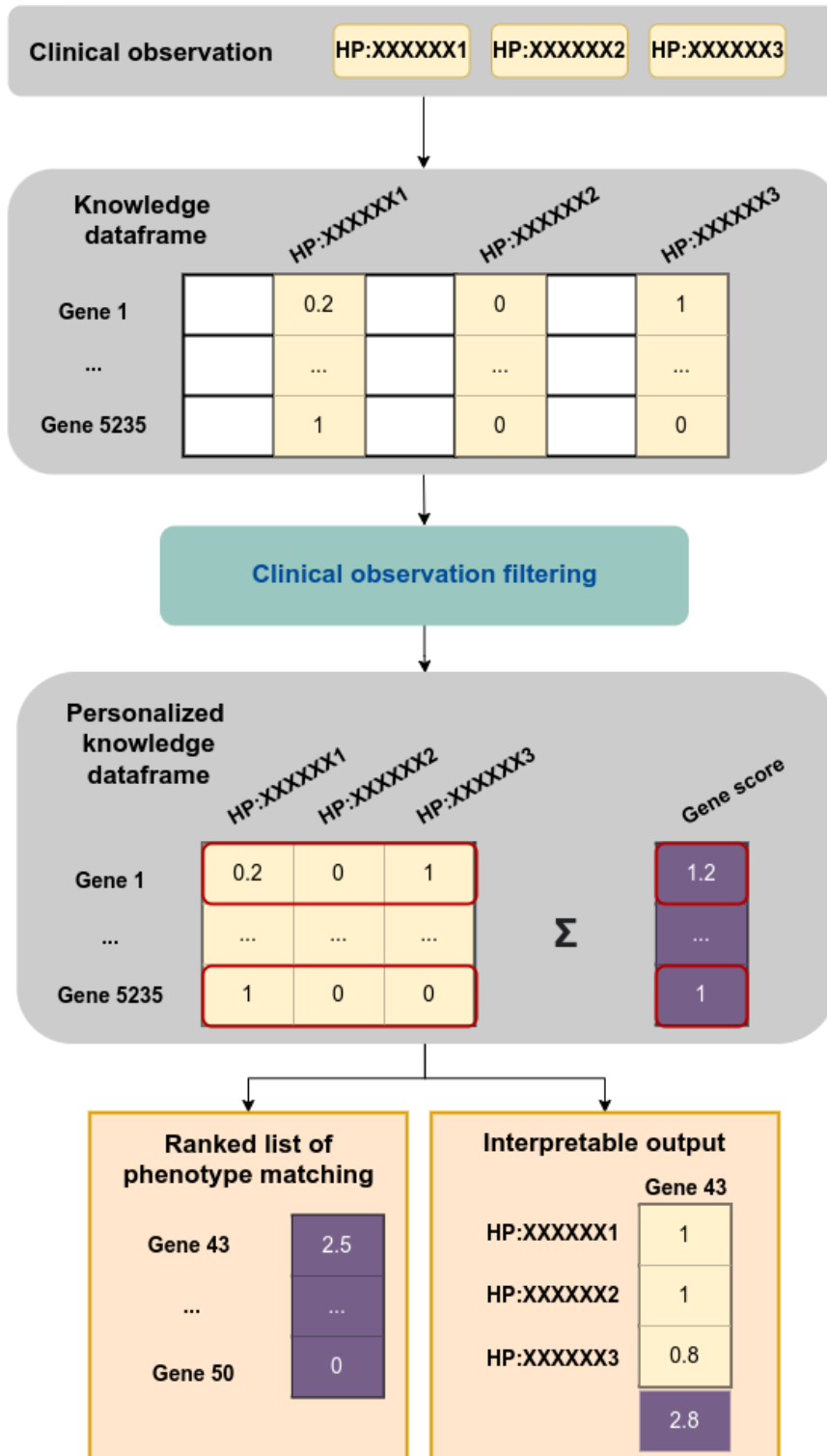
**Fig. S9. Dendrogram of hierarchical clusters of clinical observations obtained using agglomerative clustering on cohort projection in 390 groups of interacting symptoms dimension.** The count of observations per cluster was reported (n). Colors represent branches of the hierarchy.



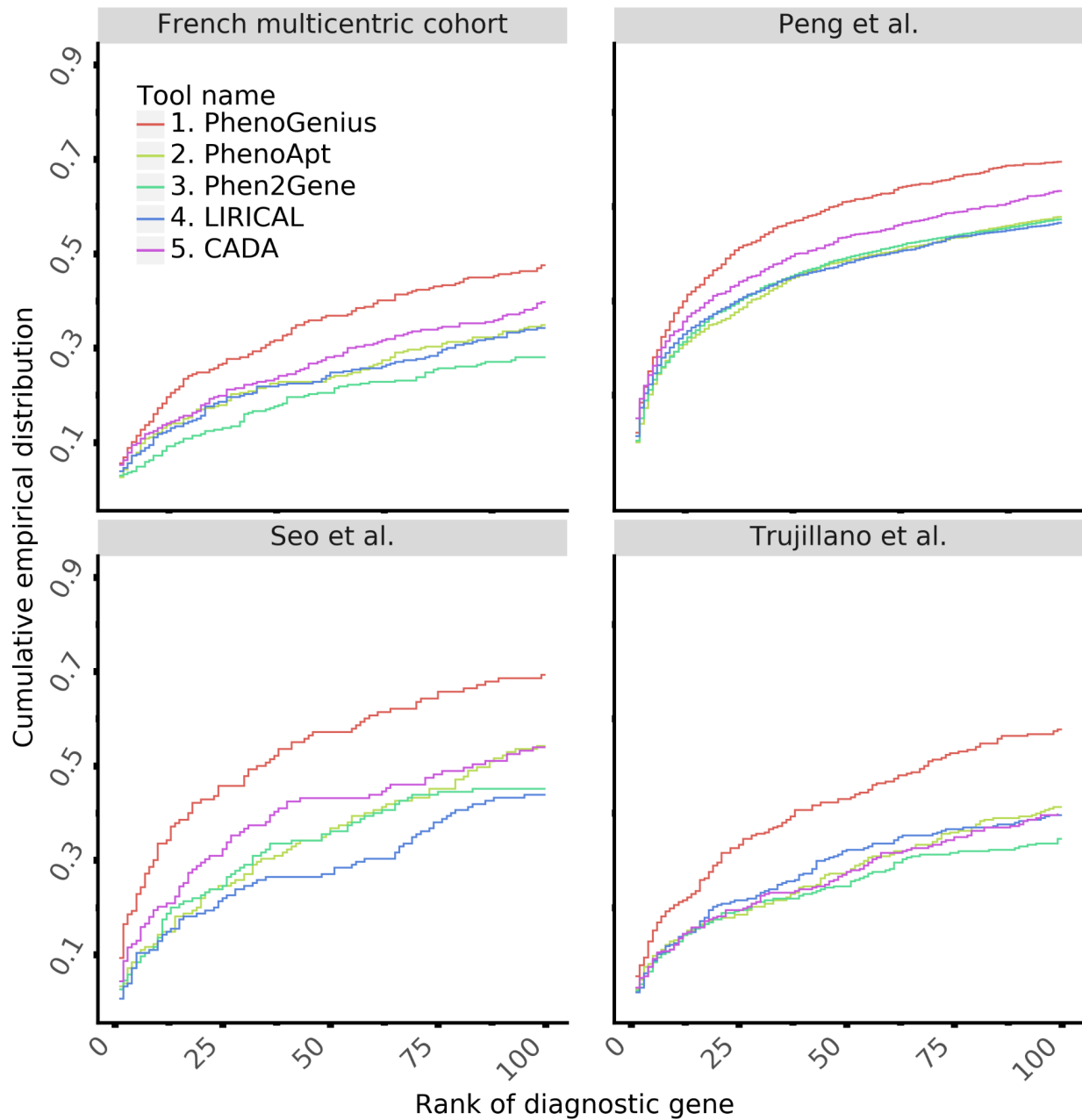
**Fig. S10. UMAP visualization of cohort's clinical descriptions projected using the 390 groups of interacting symptoms, colored and annotated by agglomerative cluster. White boxes represent clinical reports description phenotyped by twelve physicians.**



**Fig. S11. UMAP visualization of cohort's clinical descriptions projected using the 16,600 symptoms from HPO, colored by the number of symptoms. Boxes represent clinical reports description phenotyped by twelve physicians.**



**Fig. S12. Illustration of phenotype matching and gene prioritization system.** According to symptoms of a clinical description in HPO format, the knowledge dataframe is filtered. A personalized and interpretable ranking of genes is provided according to the sum of associated symptom-gene associations available.



**Fig. S13. Benchmark of a selection of state-of-the-art phenotype-driven gene prioritization per sub-group cohort.**

The fraction of cases correctly diagnosed (y-axis) is plotted against a cumulative causal gene rank.



**Table S1. Clinical data collection of 1,686 patients in our international cohort.**

ID	Description	Count of			
		Patients	Genes	Terms Total   Unique	
French multicenter cohort from PhenoGenius consortium	Gathered from CHU Grenoble Alpes, CHU de Dijon, CHU de Montpellier, CHU de Brest, and Hospices Civils de Lyon	307	220	3243	989
Seo <i>et al.</i> (1)	Unselected series of consecutive patients, clinically suspected of carrying a genetic disorder, from non-consanguineous families, who presented at the Medical Genetics Center, Asan Medical Center, Seoul, South Korea, from April 2018 to August 2019.	140	120	1135	347
Trujillano <i>et al.</i> (2)	Consecutive, unrelated patients referred by physicians from 54 countries on different continents have been included in this study. All patients with suspected Mendelian disorders were referred for diagnostic exome sequencing between January 2014 and January 2016.	298	241	2411	789
Peng <i>et al.</i> (3)	Collection of 435 descriptions from German hospitals and 506 ClinVar submissions.	941	528	6439	1814

**Table S2. Top ten recurring genes in the four groups in our cohort.**

Gene name	Count (n=1,686)	Percentage
<i>ABCC6</i>	22	1.29
<i>ANKRD11</i>	21	1.23
<i>ARID1B</i>	20	1.18
<i>NSD1</i>	18	1.06
<i>BLM</i>	16	0.94
<i>FBN1</i>	15	0.88
<i>MECP2</i>	15	0.88
<i>NF1</i>	15	0.88
<i>PTPN11</i>	14	0.82
<i>PKD1</i>	13	0.76

**Table S3. Top ten recurring HPO terms in the four groups in our cohort.**

HPO	Description	Count (n=13,228)	Percentage
HP:0001263	Global developmental delay	373	2.82
HP:0000750	Delayed speech and language development	241	1.82
HP:0001249	Intellectual disability	231	1.75
HP:0000252	Microcephaly	209	1.58
HP:0001250	Seizure	205	1.55
HP:0001252	Hypotonia	170	1.29
HP:0004322	Short stature	168	1.27
HP:0001270	Motor delay	126	0.95
HP:0001622	Premature birth	108	0.82
HP:0000486	Strabismus	102	0.77

**Table S4. Description of clinical geneticist profiles in the prospective phenotyping experiment.**

ID	Profile	Self-estimated expertise in phenotyping using HPO format (from 1 to 10)
1	Clinician	7
2	Clinician	1
3	Clinician	5
4	Resident in medical genetics	9
5	Clinician	5
6	Clinician	1
7	Clinician and laboratory specialist	1
8	Clinician	6
9	Clinician and laboratory specialist	5
10	Resident in medical genetics	6
11	Clinician	3
12	Resident in medical genetics	1

## Conclusions

The aim of my thesis project was to leverage bottlenecks in genomic medicine using bioinformatics and data science, and help pursue precision medicine adoption in healthcare. For the teamwork achievements I described in this manuscript, I was awarded the 2021 “*Sabatier d’Espeyran*” scientific prize in Montpellier, France (<https://www.ac-sciences-lettres-montpellier.fr/>). I hope I succeeded in making my contribution to spreading genomic medicine awareness in the community and providing technical solutions to improve rare diseases’ patient care.

We provided for the community *Genome Alert!* semi-automated system for genomic analysis reinterpretation that solves numerous diagnostics. Since *Genome Alert!* publication, 621 different visitors from ten countries have visited the website, 1,683 read the publication and got its first citation in July 2022. Still, *Genome Alert!* monitoring method of the ClinVar database represents a partial response to the reinterpretation task of previous genomic analysis. It doesn’t cover all processes that could improve sequencing data reinterpretation, such as upgrading the variant detection pipeline to catch additional variant types from NGS or changing reference genome <sup>30</sup>. Of course, it will never be as efficient as a complete reinterpretation performed by a clinical laboratory scientist. But I believe it provided a scalable and affordable approach to tackle this challenge. Following this method based on the data sharing community, the following steps will be to monitor other data sources of clinical knowledge, such as the challenging PubMed literature in scientific articles.

My main focus during this Ph.D. was computational phenotype analysis, as I believe the expertise gathered around this thesis was unique to tackle this challenge.

The Monarch Initiative provides tremendous collective efforts to provide the Human Phenotype Ontology and gather symptom-gene associations described in the medical literature. Studies reported methods to identify the specificity of symptoms to genes<sup>70,71</sup>. These accomplishments provided significant progress in computational phenotype analysis. However, with the current associations available in databases, half of the symptoms declared in the collected cohort of clinical descriptions were not considered. Several studies try to tackle fuzzy physicians' phenotyping using semantic interrelationships between terms<sup>66,72,73</sup> to match clinical descriptions to the knowledge database. In contrast to our approach, they rely on the HPO architecture, which is ordered according to human development, so it may not represent the interaction of symptoms in disease.

Inspired by medical inductive reasoning, we describe an original method that models symptom interaction in the global spectrum of symptom-gene associations. This model learned physicians' phenotypic patterns, successfully handled heterogeneous phenotyping, and provided an almost complete relationship between physicians' clinical descriptions and medical literature knowledge.

I hope this work will open a whole new field in computational phenotype analysis, helping to identify new genes in disease, expand their clinical spectrum and provide an easily interpretable clinical decision support system. I'll use this system in the upcoming years to tackle new bottlenecks in variant interpretation and clinically relevant group definition using the clinical phenotypes powered by symptom interaction modeling.

## References

1. Th, A., Attia, T. H. & Saeed, M. A. Next Generation Sequencing Technologies: A Short Review. *Journal of Next Generation Sequencing & Applications* **01**, (2015).
2. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
3. Nambot, S. *et al.* Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genet. Med.* (2017) doi:10.1038/gim.2017.162.
4. Adams, D. R. & Eng, C. M. Next-Generation Sequencing to Diagnose Suspected Genetic Disorders. *N. Engl. J. Med.* **379**, 1353–1362 (2018).
5. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine-Progress, Pitfalls, and Promise. *Cell* **177**, 45–57 (2019).
6. Darwin, C. *The Origin of Species*. (BEYOND BOOKS HUB, 1982).
7. Koonin, E. V. Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* **41**, 298–306 (2009).
8. Alberts, B. *Molecular Biology of the Cell*. (Garland Science, 2017).
9. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* vol. 171 737–738 (1953).
10. Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **171**, 740–741 (1953).
11. Brown, T. A. *Genomes 4*. (Garland Science, 2018).
12. Nirenberg, M. Historical review: Deciphering the genetic code—a personal account. *Trends Biochem. Sci.* **29**, 46–54 (2004).
13. Orgogozo, V., Morizot, B. & Martin, A. The differential view of genotype-phenotype

- relationships. *Front. Genet.* **6**, 179 (2015).
14. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
  15. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
  16. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
  17. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
  18. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
  19. Korf, B. R. Integration of genetics into clinical teaching in medical school education. *Genet. Med.* **4**, 33S–38S (2002).
  20. Williams, M. S. Early Lessons from the Implementation of Genomic Medicine Programs. *Annu. Rev. Genomics Hum. Genet.* **20**, 389–411 (2019).
  21. Esquivel-Sada, D. & Nguyen, M. T. Diagnosis of rare diseases under focus: impacts for Canadian patients. *J. Community Genet.* **9**, 37–50 (2018).
  22. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
  23. Muzzey, D., Evans, E. A. & Lieber, C. Understanding the Basics of NGS: From Mechanism to Variant Calling. *Curr. Genet. Med. Rep.* **3**, 158–165 (2015).
  24. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
  25. Stessman, H. A., Bernier, R. & Eichler, E. E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872–877 (2014).



26. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* **19**, 253–268 (2018).
27. Lelieveld, S. H., Veltman, J. A. & Gilissen, C. Novel bioinformatic developments for exome sequencing. *Hum. Genet.* **135**, 603–614 (2016).
28. Scholz, M. B., Lo, C.-C. & Chain, P. S. G. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* **23**, 9–15 (2012).
29. Pereira, R., Oliveira, J. & Sousa, M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *Journal of Clinical Medicine* vol. 9 132 (2020).
30. Robertson, A. J. *et al.* Re-analysis of genomic data: An overview of the mechanisms and complexities of clinical adoption. *Genet. Med.* **24**, 798–810 (2022).
31. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17**, 9–18 (2016).
32. Elliott, A. M. *et al.* Genome-wide sequencing and the clinical diagnosis of genetic disease: The CAUSES study. *HGG Adv* **3**, 100108 (2022).
33. Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.* **6**, 252ra123 (2014).
34. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).
35. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, (2016).
36. Verweij, N. *et al.* Germline mutations in *CIDEB* and protection against liver disease. *N. Engl. J. Med.* **387**, 332–344 (2022).
37. 100,000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).

38. Delude, C. M. Deep phenotyping: The details of disease. *Nature Publishing Group UK* <http://dx.doi.org/10.1038/527S14a> (2015) doi:10.1038/527S14a.
39. Rouxel, F. *et al.* Using deep-neural-network-driven facial recognition to identify distinct Kabuki syndrome 1 and 2 gestalt. *Eur. J. Hum. Genet.* **30**, 682–686 (2022).
40. Hsieh, T.-C. *et al.* GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nat. Genet.* **54**, 349–357 (2022).
41. Kohane, I. S. HEALTH CARE POLICY. Ten things we have to do to achieve precision medicine. *Science* **349**, 37–38 (2015).
42. Turro, E. *et al.* Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).
43. Stark, Z. *et al.* Australian Genomics: A Federated Model for Integrating Genomics into Healthcare. *Am. J. Hum. Genet.* **105**, 7–14 (2019).
44. Wells, C. F. *et al.* Rapid exome sequencing in critically ill infants: implementation in routine care from French regional hospital's perspective. *Eur. J. Hum. Genet.* (2022) doi:10.1038/s41431-022-01133-7.
45. Testard, Q. *et al.* Exome sequencing as a first-tier test for copy number variant detection : retrospective evaluation and prospective screening in 2418 cases. *medRxiv* 2021.10.14.21264732 (2021).
46. Pujol, P. *et al.* Predominance of Mutation and Estrogen Receptor Positivity in Unselected Breast Cancer with or Mutation. *Cancers* **14**, (2022).
47. Amendola, L. M. *et al.* Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am. J. Hum. Genet.* **98**, 1067–1076 (2016).
48. Richards, S. *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.* **17**, 405–424

- (2015).
49. Acmg Board Of Directors. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 721–722 (2017).
  50. Azzariti, D. R. & Hamosh, A. Genomic Data Sharing for Novel Mendelian Disease Gene Discovery: The Matchmaker Exchange. *Annu. Rev. Genomics Hum. Genet.* **21**, 305–326 (2020).
  51. Rehm, H. L., Harrison, S. M. & Martin, C. L. ClinVar Is a Critical Resource to Advance Variant Interpretation. *The Oncologist* vol. 22 1562–1562 (2017).
  52. Wright, C. F. *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.* **20**, 1216–1223 (2018).
  53. Berger, S. M. *et al.* Challenges of variant reinterpretation: Opinions of stakeholders and need for guidelines. *Genet. Med.* (2022) doi:10.1016/j.gim.2022.06.002.
  54. Bombard, Y. *et al.* The Responsibility to Recontact Research Participants after Reinterpretation of Genetic and Genomic Research Results. *Am. J. Hum. Genet.* **104**, 578–595 (2019).
  55. Landrum, M. J. & Kattman, B. L. ClinVar at five years: Delivering on the promise. *Hum. Mutat.* **39**, 1623–1630 (2018).
  56. Yauy, K. *et al.* Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene-phenotype reassessment in clinical routine. *Genet. Med.* **24**, 1316–1327 (2022).
  57. Baynam, G. *et al.* Phenotyping: targeting genotype’s rich cousin for diagnosis. *J. Paediatr. Child Health* **51**, 381–386 (2015).
  58. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).

59. Köhler, S., Kindle, G. & Robinson, P. N. *The Human Phenotype Ontology in 2021*. (2021).
60. Robinson, P. N. *et al.* Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *Am. J. Hum. Genet.* **107**, 403–417 (2020).
61. Evans, C. D. Computer systems in dysmorphology. *Clin. Dysmorphol.* **4**, 185–201 (1995).
62. McKusick, V. A. *OMIM(TM): Online Mendelian Inheritance in Man*. (2000).
63. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
64. Louden, D. N. MedGen: NCBI's Portal to Information on Medical Conditions with a Genetic Component. *Med. Ref. Serv. Q.* **39**, 183–191 (2020).
65. Allot, A. *et al.* LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Res.* **46**, W530–W536 (2018).
66. Köhler, S. *et al.* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85**, 457–464 (2009).
67. Shin, H. S. Reasoning processes in clinical reasoning: from the perspective of cognitive psychology. *Korean J Med Educ* **31**, 299–308 (2019).
68. Martin, A. R. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
69. Yauy, K. *et al.* Learning phenotypic patterns in genetic disease by symptom interaction modeling. *medRxiv* 2022.07.29.22278181 (2022).
70. Greene, D., NIHR BioResource, Richardson, S. & Turro, E. Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. *Am. J. Hum. Genet.* **98**, 490–499 (2016).
71. Jagadeesh, K. A. *et al.* Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet. Med.* **21**, 464–470 (2019).
72. Deng, Y., Gao, L., Wang, B. & Guo, X. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLoS One*

**10**, e0115692 (2015).

73. Shen, F. *et al.* HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology. *J. Biomed. Inform.* **96**, 103246 (2019).

## Thesis abstract

Rare diseases are individually rare but collectively frequent, with more than 7% of living adults affected by one of the 6000 currently described diseases. An estimated 72% of rare diseases are genetic in origin. Since the next generation sequencing (NGS) technology revolution, the rare diseases diagnosis bottleneck is no longer the sequencing but the analysis of the massive amount of data produced. Despite genome sequencing accessibility in clinical routine, the majority of patients suffering from rare diseases are still undiagnosed. Using bioinformatics and data science, my thesis project aimed to manage current bottlenecks of genomic medicine to improve rare disease diagnoses. This manuscript is focused on two main projects I led during this Ph.D. with SeqOne Genomics and CHU Grenoble Alpes.

First, I tackled the reinterpretation challenge of previous sequencing analysis that remained unsolved. This reinterpretation was reported manually, and the lack of human resources and automated methods made it difficult to apply in routine diagnosis. Taking advantage of the collaborative and dynamic database ClinVar of shared variant interpretation, we developed *Genome Alert!*, an open-source automated method that monitors ClinVar and monthly reassesses variant pathogenicity and symptom-gene associations. The re-interpretation of 4,929 analyses revealed 45 changes with potential clinical impact, leading to four additional diagnoses. This work represents a first large validation study of an automated sequencing data re-interpretation system that could become a standard in genomic medicine.

Lastly, I explored the clinical data computation challenge, aiming to improve the medical coding or physician's phenotyping use in genomic analysis. We report the first study focusing on phenotyping practices in clinical sequencing analysis, analyzing the records of 1,686 patients from four international groups. Despite the adoption of a common standard called Human Phenotype Ontology, we found a highly heterogeneous approach to phenotyping as regards the number and choice of symptoms, even for the same patients. This fluctuating description is a major challenge that has to be overcome to enable us to exploit the clinical data in medical records. As an illustration, less than half (43%) of declared symptom-gene associations in the cohort were covered in public databases.

Aiming to model the medical inductive reasoning that could explain the heterogeneity of phenotyping across clinical observations, we developed methods based on the association of symptoms with the same genetic disorder. Using graph algorithms and collaborative filtering, we trained a symptom interaction model that projects clinical descriptions in HPO format including 16,600 symptoms into the dimension of interacting symptoms containing 390 groups and 1,131,886 pairs of associated symptoms in diseases. This model uncovered the missing pieces of the incomplete clinical descriptions puzzle, achieving 99.8% coverage of the medical observations with knowledge in the medical literature. To evaluate its clinical relevance, we applied this symptom interaction model to phenotype-driven gene prioritization in the cohort and improved the diagnostic performance by 42 % compared to the best current competitor. This method should enable discoveries in precision medicine by standardizing clinical descriptions.

With the work described in this manuscript, I hope I succeeded in making my contribution to spreading genomic medicine awareness in the community and providing technical solutions to improve rare diseases' patient care.

1000 characters abstract :

Despite genome sequencing accessibility in clinical routine, a majority of patients suffering from rare diseases are still undiagnosed. Using bioinformatics and data science, my thesis project aimed to manage current bottlenecks of genomic medicine in patient care to improve rare disease diagnoses.

First, I tackled the reinterpretation challenge of previous sequencing analysis that remained unsolved. We developed a semi-automated method for reassessing variant pathogenicity in the ClinVar database called *Genome Alert!* that solves numerous diagnostics.

Lastly, I explored the clinical data computation challenge, aiming to improve the medical coding or physician's phenotyping use in genomic analysis. Here I described the first analysis of phenotyping practices in a clinical sequencing setting and the development of symptom interaction models in genetic diseases to provide standardized clinical descriptions and interpretable phenotype matches between symptoms and genes.



## Résumé de la thèse

Les maladies rares sont individuellement rares mais collectivement fréquentes. Plus de 7% des adultes sont affectés dans le monde par l'une des 6000 maladies actuellement décrites. 72 % des maladies rares sont d'origine génétique. Depuis l'apparition du séquençage de nouvelle génération, le diagnostic des maladies rares n'est plus limité par le séquençage en lui-même mais l'analyse des données générées par le séquençage. Malgré l'accessibilité en routine clinique du séquençage du génome, la majorité des patients souffrant de maladies rares restent sans diagnostic. Mon projet de thèse visait à résoudre des défis actuels dans l'analyse du séquençage pour améliorer le diagnostic des maladies rares. Ce manuscrit est axé sur deux principaux projets que j'ai menés au cours de ce doctorat avec l'équipe de SeqOne Genomics et le CHU Grenoble Alpes.

Premièrement, je me suis attaqué au problème de la réinterprétation des données de séquençage de patients restés sans diagnostic. Cette étape de réinterprétation est manuelle, et le manque de ressources humaines la rend difficile à réaliser en routine. Nous avons développé *Genome Alert!*, une méthode automatisée et libre qui monitore les changements dans la base de données de partage d'interprétation des variants ClinVar. Ce monitoring permet de réévaluer mensuellement et automatiquement la pathogénicité des variants et les gènes impliqués en maladies humaines. La réinterprétation de 4 929 analyses avec cette méthode a révélé 45 changements ayant un impact clinique potentiel et a conduit à quatre diagnostics supplémentaires. Ce travail représente la première validation à grande échelle d'un

système automatisé de réinterprétation des données de séquençage qui pourrait devenir un standard en médecine génomique.

En seconde partie, j'ai exploré le défi de la numérisation des données cliniques, avec pour objectif d'améliorer l'utilisation du phénotypage (ou codage médical) des cliniciens dans l'analyse génomique. Nous rapportons la première étude axée sur les pratiques de phénotypage, en analysant 1 686 descriptions de patients provenant de quatre groupes internationaux. Malgré l'adoption d'une norme commune appelée Human Phenotype Ontology, nous avons constaté une approche très hétérogène du phénotypage en ce qui concerne le nombre et le choix des symptômes, et ce même pour les mêmes patients. Cette description fluctuante est un défi majeur qui doit être surmonté pour nous permettre d'exploiter les données cliniques des dossiers médicaux. En effet, moins de la moitié (43%) des associations symptôme-gène déclarées dans la cohorte étaient retrouvées dans les bases de données publiques. Dans le but de modéliser ce raisonnement médical inductif qui pourrait expliquer l'hétérogénéité du phénotypage entre les observations cliniques, nous avons développé des méthodes basées sur l'association conjointe de symptômes au sein des maladies génétiques.

À l'aide d'algorithmes graphes, nous avons entraîné un modèle d'interaction des symptômes en maladies génétiques qui projette les descriptions cliniques en format HPO (16,600 symptômes) dans la dimension des symptômes en interaction contenant 390 groupes et 1 131 886 paires de symptômes. Pour évaluer la pertinence clinique de ce modèle, nous l'avons utilisé comme système de

priorisation de gènes en fonction du phénotype et avons amélioré les performances de priorisation de 42 % par rapport au meilleur concurrent actuel. Ce modèle devrait permettre de nouvelles découvertes en médecine de précision par sa capacité à exploiter des descriptions cliniques hétérogènes.

Au travers ce travail de thèse, j'espère avoir réussi à apporter ma pierre à l'édifice pour sensibiliser à la médecine génomique dans la communauté médicale et fournir des solutions techniques pour améliorer la prise en charge des patients atteints de maladies rares.

Résumé de 1000 caractères :

Malgré l'accessibilité en routine du séquençage du génome, la majorité des patients souffrant de maladies rares restent sans diagnostic. Mon projet de thèse avec SeqOne Genomics et le CHU Grenoble Alpes visait à résoudre des défis de l'analyse du séquençage pour améliorer le diagnostic des maladies rares.

Je me suis attaqué au problème de la réinterprétation des données de séquençage de patients restés sans diagnostic. Nous avons mis au point la première méthode semi-automatique de réévaluation de la pathogénicité des variants dans la base de données ClinVar, appelée *Genome Alert!*, qui permet de résoudre de nombreux diagnostics.

J'ai aussi exploré le défi de la numérisation des données cliniques, pour améliorer l'utilisation du phénotypage (ou codage médical) dans l'analyse génomique. J'y décris la première analyse des pratiques de phénotypage et le développement de

modèles d'interaction des symptômes dans les maladies génétiques afin d'obtenir des descriptions cliniques standardisées.