Understanding the Code of Life: Holistic Conceptual Modeling of the Genome

Diciembre de 2022

Alberto García Simón

Directores:

Prof. Dr. Óscar Pastor López
Prof. Dr. Juan Carlos Casamayor Rodenas

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Esta tesis ha sido redactada para la obtención del título de Doctor en Ingeniería informática por la Universitat Politècnica de València y defendida el {dia} de diciembre de 2022.

Autor:

Alberto García Simón

Directores:

Prof. Óscar Pastor López

Prof. Juan Carlos Casamayor Rodenas

Comité de evaluadores externos:

Prof. Johann Eder

Prof. Manuel Pérez Alonso

Prof. Jolita Ralyté

Tribunal de defensa:

Presidente: Prof. Jolita Ralyté

Secretario: Prof. José Ignacio Panach Navarrete

Vocal: Joao Falcao E Cunha

Acknowledgements

En primer lugar, quiero agradecer a mis padres por todo su apoyo incondicional durante estos años y por haberme ayudado a llegar hasta aquí. Gracias, esta tesis es tan mía como vuestra. También al resto de mi familia, especialmente a Mari. Hace ya mucho tiempo, me convenciste para perseguir una meta que ya tenía descartada. Por ello, una parte de esta tesis también es tuya.

A ti, mi mejor amiga, mi compañera de vida. Gracias por tu tiempo, por tu paciencia y por todo tu amor. Sin ti, nada de esto habría sido posible.

A mis amigos, especialmente a Alfex, Xion, Pablo, Andrea, Tere, Angie y Mar. Gracias, sois una parte central de mi vida.

A mis directores de tesis, Oscar y Juan Carlos, y a Giancarlo Guizzardi. Gracias por vuestro continuo soporte, vuestras enseñanzas y por no haber dudado de mi en ningún momento. Me habéis guiado a través de este viaje de manera brillante. *Giancarlo, you are not only a great researcher, but an amazing person.*

A Anna bernasconi. Eres una de las mejores investigadoras con las que he trabajado, me inspiras a esforzarme y dar lo mejor de mí mismo.

Besides Oscar, Juan Carlos, and Giancarlo, I would like to thank the rest of my thesis committee: Jolita Ralyté, José Ignacio Panach Navarrete, and Joao Falcao E Cunha, as well as the reviewers Prof. Johann Eder, Prof. Manuel Perez Alonso, Prof. Jolita Ralyté. Your participation in the last stage of the thesis has been a great honour.

A José Marín, gracias por todo el tiempo que has dedicado desinteresadamente a revisar y mejorar mi trabajo.

A todos mis compañeros del PROS, en especial a José, Ana L., Mireia, Ana P., René y Rosa. Gracias por todo vuestro tiempo, consejos, paciencia y por hacer de nuestro trabajo una actividad agradable que en no pocas ocasiones me ha alegrado el día.

Las fructíferas discusiones y los resultados derivados de los proyectos INNEST2021/57, MICIN/AEI/10.13039/501100011033, PID2021-123824OB-I00, CIPROM/2021/023 y PDC2021-121243-I00 han contribuido en gran medida a la calidad final de este tesis.

A Víctor

Abstract

Over the last few decades, advances in sequencing technology have produced significant amounts of genomic data, which has revolutionised our understanding of biology. However, the amount of data generated has far exceeded our ability to interpret it.

Deciphering the code of life is a grand challenge. Despite our progress, our understanding of it remains minimal, and we are just beginning to uncover its full potential, for instance, in areas such as precision medicine or pharmacogenomics.

The main objective of this thesis is to advance our understanding of life by proposing a holistic approach, using a model-based approach, consisting of three artifacts: i) a conceptual schema of the genome, ii) a method for its application in the real-world, and iii) the use of foundational ontologies to represent domain knowledge in a more unambiguous and explicit way. The first two contributions have been validated by implementing genome information systems based on conceptual models. The third contribution has been validated by empirical experiments assessing whether using foundational ontologies leads to a better understanding of the genomic domain.

The artifacts generated offer significant benefits. First, more efficient data management processes were produced, leading to better knowledge extraction processes. Second, a better understanding and communication of the domain was achieved.

Resumen

En las últimas décadas, los avances en la tecnología de secuenciación han producido cantidades significativas de datos genómicos, hecho que ha revolucionado nuestra comprensión de la biología. Sin embargo, la cantidad de datos generados ha superado con creces nuestra capacidad para interpretarlos.

Descifrar el código de la vida es un gran reto. A pesar de los numerosos avances realizados, nuestra comprensión del mismo sigue siendo mínima, y apenas estamos empezando a descubrir todo su potencial, por ejemplo, en áreas como la medicina de precisión o la farmacogenómica.

El objetivo principal de esta tesis es avanzar en nuestra comprensión de la vida proponiendo una aproximación holística mediante un enfoque basado en modelos que consta de tres artefactos: i) un esquema conceptual del genoma, ii) un método para su aplicación en el mundo real, y iii) el uso de ontologías fundacionales para representar el conocimiento del dominio de una forma más precisa y explícita. Las dos primeras contribuciones se han validado mediante la implementación de sistemas de información genómicos basados en modelos conceptuales. La tercera contribución se ha validado mediante experimentos empíricos que han evaluado si el uso de ontologías fundacionales conduce a una mejor comprensión del dominio genómico.

Los artefactos generados ofrecen importantes beneficios. En primer lugar, se han generado procesos de gestión de datos más eficientes, lo que ha permitido mejorar los procesos de extracción de conocimientos. En segundo lugar, se ha logrado una mejor comprensión y comunicación del dominio.

Resum

En les últimes dècades, els avanços en la tecnologia de seqüenciació han produït quantitats significatives de dades genòmiques, fet que ha revolucionat la nostra comprensió de la biologia. No obstant això, la quantitat de dades generades ha superat amb escreix la nostra capacitat per a interpretar-los.

Desxifrar el codi de la vida és un gran repte. Malgrat els nombrosos avanços realitzats, la nostra comprensió del mateix continua sent mínima, i a penes estem començant a descobrir tot el seu potencial, per exemple, en àrees com la medicina de precisió o la farmacogenómica.

L'objectiu principal d'aquesta tesi és avançar en la nostra comprensió de la vida proposant una aproximació holística mitjançant un enfocament basat en models que consta de tres artefactes: i) un esquema conceptual del genoma, ii) un mètode per a la seua aplicació en el món real, i iii) l'ús d'ontologies fundacionals per a representar el coneixement del domini d'una forma més precisa i explícita. Les dues primeres contribucions s'han validat mitjançant la implementació de sistemes d'informació genòmics basats en models conceptuals. La tercera contribució s'ha validat mitjançant experiments empírics que han avaluat si l'ús d'ontologies fundacionals condueix a una millor comprensió del domini genòmic.

Els artefactes generats ofereixen importants beneficis. En primer lloc, s'han generat processos de gestió de dades més eficients, la qual cosa ha permés millorar els processos d'extracció de coneixements. En segon lloc, s'ha aconseguit una millor comprensió i comunicació del domini.

Contents

Acknowledgement iii

Abstract v

Contents xi

I Preface 1

1 Introduction 3

1.1 Motivation 3

1.2 Problem Statement. 9

1.3 Research Methodology 11

1.4 Objectives and Research Questions 12

1.5 Thesis Summary 15

II Main 17

2 Problem Investigation 19

2.1 Evolution of Genomics 20

2.2 Genomics Data Management Status 34
2.3 Genomics Data Management Strategies 35
2.4 Conclusions 39

3 Treatment Design 41
3.1 The Conceptual Schema of the Human Genome 42
3.2 The Conceptual Schema of the Citrus Genome 54
3.3 The Conceptual Schema of the Genome 59
3.4 The ISGE Method 81
3.5 The Ontological Unpacking 85
3.6 Conclusions 94

4 Treatment Validation 97
4.1 The Delfos Oracle Platform 98
4.2 The CitrusGenome Platform 112
4.3 Ontological Unpacking Assessment 133
4.4 Conclusions 144

III Results 147

5 Conclusions 149
5.1 Answer to Research Questions 150

6 Thesis Impact 155

7 Future Work 161
7.1 Final considerations 166

Bibliography 169

A The Complete List of Publications 191

List of Figures

1.1 Oxidative phosphorylation pathway 4

1.2 NADH:ubiquinone oxidoreductase reaction 5

1.3 Knowledge and limitations 6

1.4 Computer science vs genomics 8

1.5 Thesis' Design Cycle 11

2.1 Summary of genomics milestones 21

2.2 DNA and RNA structure 22

2.3 Central dogma of biology 23

2.4 Semiconservative replication of the DNA. 23

2.5 Sanger process 25

2.6 Cost of sequencing a human genome 27

2.7 Reformulated central dogma 29

2.8 Cost of sequencing a human genome 30

2.9 Second-generation DNA technique 32

2.10 Third-generation DNA technique 33

3.1 Achieving a more agnostic approach 45

3.2 Achieving the representation of assemblies 47

3.3 Achieving the new representation of variations 49

3.4 Achieving a first approach to model variation effects 51

3.5 Achieving a better representation of gene expression processes . 53

3.6 Conceptual Schema of the Citrus Genome 56

3.7 The structural view of the CSG 62

3.8 The transcription view of the CSG 63

3.9 The variation view of the CSG 64

3.10 Achieving an improved representation of proteins 69

3.11 Achieving a better representation of the clinical significance . . 74

3.12 Achieving an improved representation of biological entities . . . 77

3.13 Achieving a better representation of the roles of biological entities 79

3.14 The difference between a view and a conceptual view 82

3.15 The Ontological unpacking 85

3.16 Instance of the Ontological unpacking 90

4.1 Instantiation of the ISGE method for the human case 101

4.2 Instantiation of the ISGE method for the citrus case 115

4.3 GenomIUm phases. 123

4.4 Subject's previous experience 135

List of Tables

3.1 Example of helix structures 71

3.2 Example of clinical significances 75

3.3 The first template of the ISGE method 83

3.4 The second template of the ISGE method 84

3.5 The third template of the ISGE method 84

3.6 OntoUML class stereoypes 88

3.7 OntoUML relationship stereoypes 89

4.1 Select phase of the ISGE method for obtaining the CSGD . . . 104

4.2 Concepts considered to generate the CSGD 105

4.3 Example of Visualization patterns in Sibila 108

4.4 Summary of variations analysed using Delfos 110

4.5 Select phase of the ISGE method for obtaining the CSGD . . . 119

4.6 Inconsistencies in terms of knowledge gaps for the CSGD 120

4.7 Proposal to improve the CSG 121

4.8 Concepts considered to generate the CSGC 122
4.9 Example of visualization patterns in CitrusGenome 124
4.10 CitrusGenome validation participants 126
4.11 Usability problems identified in CitrusGenome 130
4.12 Questions posed to subjects. 136
4.13 Experiment procedure 137
4.14 Data analysis results for effectiveness metrics 139
4.15 Data analysis results for efficiency metrics 139
4.16 Data analysis results for satisfaction metrics 140

6.1 Summary of Relevant Publications 156
6.2 Publications boing considered for publications. 157

A.1 Detailed list of Publications. 193

Part I

Preface

Chapter 1

Introduction

1.1 Motivation

What is life? Despite being an age-old question, there is no consensus on the exact definition of life [1]–[4]. Both scientists and philosophers have offered multiple interpretations of what life is over time. Indeed, it remains a mystery that we are trying to unravel, and it is all but impossible to answer this question from an analytical perspective.

> *Life is something edible, lovable, or lethal.*
>
> **James E. Lovelock**

There is a statement that we can strongly assert: "life is complex". It is a vast network of processes, reactions, activities, and events with almost infinite variability. Life adapts to the context and its circumstances. We can find organisms that are capable of surviving in the absence of water, in extreme temperatures, being affected by enormous amounts of radiation, etc. [5]. We also find organisms that are "at the edge of life", such as viruses, as they defy the most basic principles of life [6], [7].

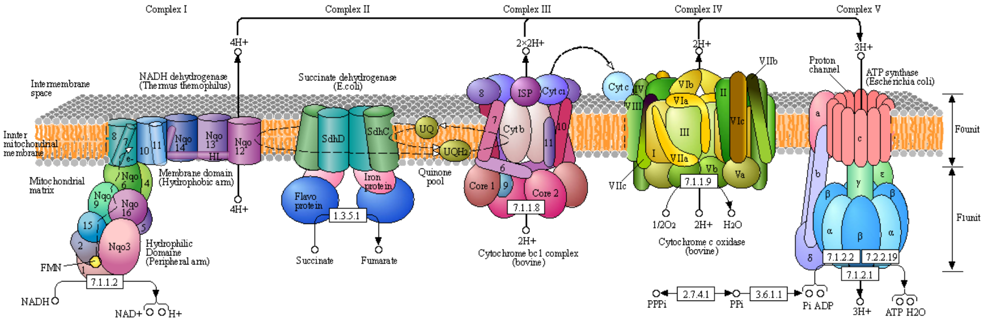

Figure 1.1: Oxidative phosphorylation pathway. ©Kanehisa Laboratories. Source: The Kyoto Encyclopedia of Genes and Genomes (KEGG).

Although life has created this high degree of heterogeneity, there are two characteristics that all living organisms share. The first characteristic is that every organism is made up of cells, and these cells contain genetic information. Genetic information can be encoded in DNA, as in humans, or in RNA, as in viruses. In addition, genetic information is structured in chromosomes, which have different structures such as linear chromosomes in the case of humans or circular chromosomes in the case of some bacteria. This genetic information is what defines organisms, shaping both their physical appearance and their behaviour. Understanding how this genetic information works is one of the main challenges humankind is currently facing [8].

The study and understanding of genetic information (i.e., the genome) is called genomics and is a field within molecular biology [9]. Genomics includes the study of the structure, function, evolution, editing, and mapping of the genome. Genomics offers a more global and holistic perspective when compared to genetics, which focuses on the study of genes individually. There are several fields of study in genomics, such as functional genomics (i.e., the study of interactions between genome components) [10], structural genomics (i.e., the study of the three-dimensional structure of genome components) [11], or epigenomics (i.e., the study of the interaction between the genome and the environment) [12].

The second characteristic shared by all living things is that life is “just” a set of chemical reactions in its most elementary stages. For instance, the synthesis of Adenosine Triphosphate (ATP), which is the fuel that provides energy to living cells, is synthesised in multiple ways. One of these is the oxida-

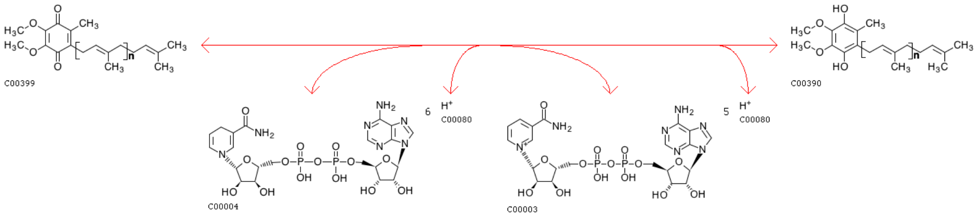

Figure 1.2: The NADH:ubiquinone oxidoreductase reaction as part of the Oxidative phosphorylation pathway. C00399 refers to Ubiquinone, C00004 refers to NADH, C00080 refers to the proton, C00390 refers to Ubiquinol, and C00003 refers to NAD+.

tive phosphorylation process[1], which consists of a set of 21 precisely described chemical reactions (Figure 1.1 shows a detailed representation of the chemical reactions that occur during oxidative phosphorylation). By way of illustration, the leftmost reaction of Fig. 1.1, catalysed by the *NADH:ubiquinone reductase* enzyme, whose EC number is EC:7.1.1.2[2], consists of the synthesis of three compounds: Ubiquinol (a reduced form of the coenzyme Q_{10}), NAD+ (an oxidated form of Nicotinamide adenine dinucleotide), and H+ (a proton). The complete reaction can be seen in Fig. 1.2. The author of this thesis is aware of the specificity and complexity of this description of the NADH:ubiquinone oxidoreductase reaction. However, we consider that it is appropriate to introduce such a level of detail to illustrate the complexity of the domain under study, with a strong, essential biomolecular and chemical dimension.

As a consequence of the two characteristics explained above (i.e.,that *every organism is made up of cells, which contain genetic information* and that *life is nothing but a set of chemical reactions*), the following conclusions can be inferred: i) the phenotype (i.e., the observable traits, the developmental processes, and the behaviour of an organism) is a consequence of the genotype (i.e., the genetic information encoded in cells), ii) reliable associations between the two can be established, and iii) these associations are somehow guided by the elementary chemical reactions of the organism [13], [14]. Although many significant advances have been achieved in this regard, the vast majority of the knowledge remains unknown.

[1] `https://www.genome.jp/kegg-bin/show_pathway?map00190`

[2] This notation is called the Enzyme Commission number, and it is used to classify enzymes based on the chemical reactions that they catalyse. In this case, the first number indicates that the enzyme is a Translocase, meaning that it catalyses the movement of ions or molecules across membranes or their separation within membranes (i.e., a membrane transport protein). The second number indicates that the enzyme transfers hydrons (i.e., a cationic form of Hydrogen, like protons). The third number indicates that the movement is linked to oxidoreductase reactions. The fourth number identifies the specific enzyme.

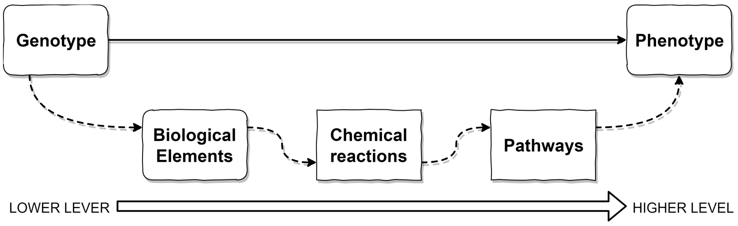

Figure 1.3: A representation of the current knowledge and its limitations.

For instance, we can establish a reliable association between the DNA (i.e., the genotype) and a phenotype, as in *cri du chat* syndrome. *Cri du chat syndrome* (French for cat cry) is a rare genetic disorder that produces several conditions, such as problems in the larynx and nervous system (causing babies to make a cat-like sound when they cry), cognitive and motor disabilities, a small head, or unusual facial features, among others. This syndrome is caused by a partial deletion of the small arm of chromosome 5. However, this association is a black box that fails to answer fundamental questions about the specific reason for each of these conditions. This problem is a common concern among geneticists and clinicians, and a significant amount of research is aimed at identifying the components of this black box. Beyond rare disease such as this one, more common diseases, like cancer are also related to this problem. Tens of thousands of variants are associated with cancer, but several questions, such as intratumoral heterogeneity [15], remain unresolved. As a result, we are unable to effectively detect, fix, and even correct cancer [16].

Going deeper into the discussion, phenotypes do not only include diseases but encompass virtually any aspect of a living being. Physical characteristics such as eye colour or height are determined by the genotype (as well as its interaction with the environment). Other non-tangible aspects, such as envy or anger (i.e., our personality), are shaped by the information encoded in our cells.

Fig 1.3 shows our current understanding with respect to the three conclusions reported above (i.e., that *the phenotype is a consequence of the genotype*, that *associations can be established between them*, and that *these associations are driven by the elementary chemical reactions of the organism*). We have succeeded in establishing such associations based on a set of standards and guidelines that are considered to be relatively reliable [17], between the genotype and the phenotype through multiple types of studies such as genotype-phenotype association studies or genome-wide association studies. However, these links result from complex interactions between biological elements, chemical reactions, and biological pathways. The genotype is responsible for the identification (e.g., genes) and creation (e.g., proteins) of biological elements. They interact with each other and with other elements to perform chemical reactions, which are chained together to create pathways.

Despite having successfully established genotype-phenotype associations, many limitations remain. There are several biological elements that have not yet been identified or whose functional characterisation is unknown. For instance, according to its official statistics, UniProt[3] (a database containing protein sequences and functional information) contains 214,406,399 entries, of which only 0.8% of them have evidence of their existence at the protein level, and almost 70% of them are predicted. Moreover, only 564,638 of the proteins stored in UniProt have been manually curated. The existence of biological entities whose functionality is unknown implies that the chemical reactions occurring in an organism are not fully understood, indicating crucial gaps in knowledge. Although we can model the steps of some of the existing biological pathways, we cannot accurately determine their inner working characteristics if the chemical reactions that compose pathways are not fully understood.

In short, the links between biological pathways and phenotypes become obscure and difficult to establish because of the reasons discussed above. We -humans - lack a holistic and global perspective that would allow us to integrate this knowledge to truly understand life, as there is a fuzzy cloud of knowledge waiting to be unravelled.

> “ *A living organism is a computer or machine made up of genetic circuits in which DNA is the software that can be hacked.*
>
> **Drew Endy** ”

In the words of Drew Endy, there is a clear analogy between living organisms and computers. In computers, the working language is the binary code, in which there are two atomic elements that are used to build up complex systems: zero and one. The binary code is the basis for creating language-specific primitives that allows us to increase the level of abstraction. We have created many programming languages that use their primitives to create reusable blocks that perform specific tasks, called functions. Functions are combined to create more complex programs whose execution yield a given result.

[3] https://www.uniprot.org

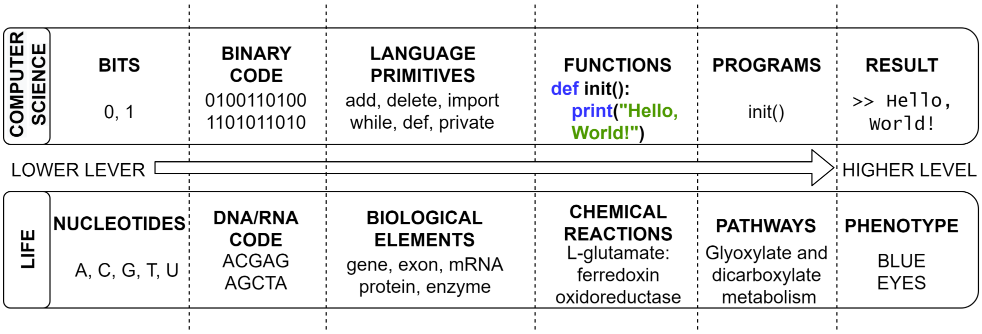

Figure 1.4: An analogy between the elements of computer science and genomics.

In life, the working language is the genetic code that is composed of DNA[4], in which there are four atomic elements[5] that are the basic components of life: adenine (A), cytosine (C), guanine (G), and thymine (T). Unlike computers, there is only one language that uses the DNA/RNA code: life. The genetic code itself can be seen as the basis for the different biological components of the genome, such as genes, proteins, enzymes, non-coding RNAs, etc. These elements represent a higher level of abstraction, similarly to the primitives of a programming language. For instance, genes abstract a DNA sequence into an atomic, biological element with a given functionality[6]. These components can perform specific activities or chemical reactions, which are generic, reusable, and function-specific (just like functions in computers). A set of chemical reactions occurring in a specific order is defined as a biological pathway, the result of which is a specific phenotype. The execution of a computer program yields a result, just as the execution of biological pathways expresses a phenotype. This analogy, explored in previous work [24], has motivated us to contribute to the laudable task of understanding life using a computer science perspective.

> *If you can write DNA, you are no longer limited to 'what is' but to what you could make.*
>
> **Drew Endy**

[4] In biology, the DNA is a polymer molecule that is made up of a repetition of monomers.

[5] In biology, these atomic elements are called monomers.

[6] We have simplified the notion of gene by way of illustration. Discussions regarding its underlying concept and implications are recurrent throughout history [18], [19], and they still exist [20]–[23].

Again, Drew Endy's words are of great interest. Although trying to edit what is not yet fully understood may seem risky, this is what has happened. Thanks to the advance in DNA editing techniques such as CRISP-R [25], we have the ability to write and edit the genetic code. These advances have already reported significant progress, such as decreasing the time that it takes to create mouse models of human diseases, studying genes, or being able to modify multiple genes in cells at once [26]. However, ethical and safety concerns have been raised regarding the use of these technologies [27]. While ethical concerns are inherent in human nature, safety concerns arise from our lack of knowledge about genomics and life. Some of the questions that arise with respect to DNA editing techniques are: Will these modifications introduce undesired changes in other parts of the genome? What are the consequences of these modifications for future generations? To answer them, we must first have a complete understanding of genomics and life.

Our vision is inspired by the analogy between the inner workings of computer science and genomics. We seek to unravel the hidden knowledge of genomics so that we can understand all of its aspects. This could allow us, on the one hand, to answer philosophical questions such as who we are, how to prevent diseases, or why life is the way it is, and, on the other hand, to have the ability to edit the genome to provide efficient, effective, and safe precision medicine. **We believe that overcoming this challenge will be one of the greatest successes that humankind can achieve.**

1.2 Problem Statement

As mentioned above, deciphering life is one of the most complex challenges we can strive to achieve. Our understanding of how genomics works and the internal processes that make living beings who they are is very limited, but our potential will be limitless if we succeed in understanding life and how it works. However, we must transform such a great challenge into more concrete ideas.

How to transform this great challenge into more concrete ideas? We decided to work with domain experts to understand their problems and limitations and then develop solutions that allow them to generate deeper and enhanced knowledge. As a result of our interaction with domain experts, we have considered two dimensions, namely, *domain conceptualisation* and *genomics data management.*

As for *domain conceptualisation*, there is no clear ontological basis for defining the most relevant concepts, and domain knowledge is continuously evolving. The same concept can be represented in different and ambiguous ways (for example, the concepts of mutation, polymorphism, and variation refer to changes in the DNA sequence), and different concepts can be identified with the same term (for example, the concept of gene can refer to a functional protein-coding unit or to the composition of non-adjacent DNA sequences). In terms of *genomics data management*, the amount of available data is enormous, scattered across hundreds of data sources, and presents a high degree of heterogeneity. These problems are captured by the term "genomic data chaos", which is described in more detail in Section 2.

Conceptual modelling allows us to consider both dimensions: conceptual and practical. Our research group had been using conceptual models to understand genomics for many years. In 2008, the director of this thesis, **Oscar Pastor**, published an article entitled "Conceptual Modelling Meets the Human Genome" [28]. He proposed the use of Conceptual Modelling as a sound and robust approach to accurately understand the human genome from a holistic perspective.

> “ *As a precise interpretation of the Human Genome would be much easier if the underlying model were known, Conceptual Modelling can provide new ways of facing that problem in order to obtain new and better strategies and solutions.*
>
> **Oscar Pastor** ”

Our research group began to conceptualise the human genome. As a result, a first version of the Conceptual Schema of the Human Genome (CSHG) was created. Over time, the CSHG has been updated as additional knowledge in genomics emerged. The CSHG underwent two major updates, in 2012 [29] and 2016 [30]. More details on the CSHG and these two updates are found below in Section 2.3.

The CSHG has been applied in various academic scenarios, including the construction of Genomic Information Systems [31]–[36], the integration of genomic data [29], [37], the improvement of communication and knowledge transfer [30], variant classification [38]–[40], or the improvement of the design of User Interfaces of genomic tools [41]–[43] as some of the most relevant scenarios.

After using the CSHG in all the above mentioned scenarios, we wanted to see its in real-world application in precision medicine. However, we had to update and extend some dimensions of the CSHG before we could use it in real-world scenarios. First, its extensive use in academic scenarios showed that certain aspects required improvement. Second, the goals and needs of real-world use cases, such as precision medicine, are more complex when compared to academic scenarios. Third, the last major update of the CSHG was conducted in 2016 [30], and genomics has continued to evolve.

Are there genomic entities that should be added, updated, or removed from the CSHG? Which areas of study should we address in more detail? How should the update and extension of the CSHG be performed? Which concepts will need to be reviewed and updated more frequently? Our eagerness to answer all these questions precisely marks the beginning of this Ph.D. thesis.

1.3 Research Methodology

This research followed the Design Science methodology by Roel Wieringa [44], which consists of the design and investigation of an artifact in a context. In this thesis, we generated a **Conceptual Schema of the Genome** (artifact) to improve domain conceptualisation and data management in **genomics** (context). A design cycle, consisting of three phases, guided our work (see Fig.1.5):

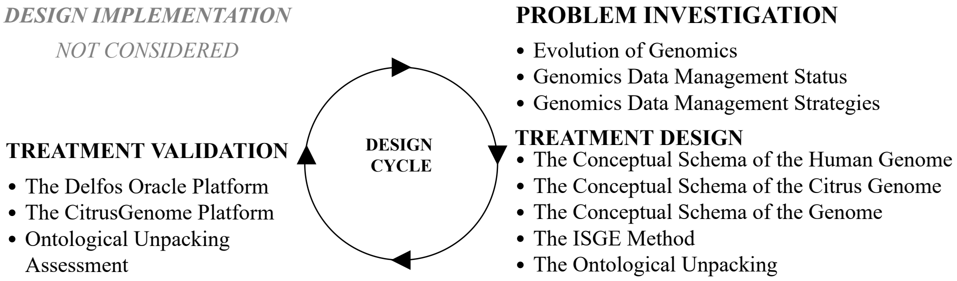
DESIGN IMPLEMENTATION
NOT CONSIDERED
PROBLEM INVESTIGATION
• Evolution of Genomics
• Genomics Data Management Status
• Genomics Data Management Strategies
TREATMENT DESIGN
• The Conceptual Schema of the Human Genome
• The Conceptual Schema of the Citrus Genome
• The Conceptual Schema of the Genome
• The ISGE Method
• The Ontological Unpacking
TREATMENT VALIDATION
• The Delfos Oracle Platform
• The CitrusGenome Platform
• Ontological Unpacking Assessment
DESIGN CYCLE

Figure 1.5: The Design Cycle defined for this thesis.

1. **Problem Investigation**: The problem context is described.
2. **Treatment Design**: The proposed treatment is provided (i.e., an artifact in a context).

3. **Treatment Validation**: The provided treatment is validated in a specific context.

1.4 Objectives and Research Questions

In this section, we identify the goals of this thesis and formulate the corresponding research questions to be answered. Our research starts with the definition of our first goal:

■ **G1** - *Study the main problems in genomics data management and how they are mitigated.*

G1 led us to the following research questions:

RQ1 - Which problems arise when working with genomics data?
RQ2 - What existing approaches can be used to mitigate the identified problems?

Our next goal is to improve the existing problems when managing genomics data. We will use conceptual modelling for this purpose.

■ **G2** - *Generate conceptual modelling artifacts to improve genomics data management.*

We have divided G2 into five sub-goals that are more specific and tangible. The first subgoal is to update the CSHG such that it can be used in real-world applications. The CSHG is a valuable asset tested in many academic contexts. However, it must be updated and extended to be helpful in real-world use cases regarding precision medicine. Thus, the first subgoal of G2 is defined:

– **G2.1** - *Extend the current version of the CSHG to be used in real-world use cases.*

G2.1 led us to the following research question:

RQ3 - Why/What/How to extend and update the CSHG?

After updating and extending the CSHG to be potentially used in real-world use cases related to precision medicine, we will broaden the scope of our research. At this point, we will have used conceptual modelling techniques to represent human genomics, giving rise to the following question: can conceptual modelling be applied in other genomic contexts? We will explore the conceptualisation of the genome of other species to answer this question, which defines the second subgoal of G2:

– **G2.2** - *Explore the conceptualisation of the genome for another species rather than humans.*

G2.2 led us to the following research question:

RQ4 - Why/What/How to generate a conceptual schema of the genome for non-human species?

To answer this research question, we will conceptualise the inner workings of another species' genome, namely, citrus. As a result, the Conceptual Schema of the Citrus Genome (CSCG) will be developed.

After updating and extending the CSHG and creating the CSCG, we will have two conceptual schemes representing two instances of the genome (i.e., the CSHG for humans and the CSCG for citrus). The characterization of the genome of different species is the same at its most elemental level, meaning that it could be plausible to generate a conceptual schema of the genome that is species-independent. Such a possibility generated the third subgoal of G2:

– **G2.3** - *Generate a conceptual schema that is species-independent.*

G2.3 led us to the following research question:

RQ5 - Why/What/How to generate a conceptual schema of the genome that is species-independent?

We will create a generic-enough artifact to be used regardless of the species under study, called the Conceptual Schema of the Genome (i.e., the CSG). The CSG will cover the most relevant aspects of the genome regardless of the species; such a broad conceptualization will lead to a sizeable conceptual schema.

Then, we will facilitate the adoption of the CSG to work with genomics use cases. Since such use cases are diverse and tend to present their specific idiosyncrasies and particularities, a method for generating subschemes containing a subset of relevant concepts of the CSG is a desirable approach to facilitate the adoption of the CSG. This generated the fourth subgoal of G2:

- **G2.4** - *Provide a method to facilitate the adoption of the CSG to work with genomics use cases..*

  G2.4 led us to the following research question:

  RQ6 - Why/What/How to create a method to generate subschemes of the Conceptual Schema of the Genome?

  Executing these conceptual modelling efforts showed us that traditional conceptual modelling artifacts are insufficient for capturing genomics particularities correctly. These limitations force us to search for other ways for better representing genomics, which led us to the fifth subgoal of G2:

- **G2.5** - *Identify another artifact for better representing genomics.*

  Ontology-driven conceptual modelling could be an excellent solution used to overcome the limitations of traditional modelLing as it considers additional semantics when modelling complex domains. Ontology-driven conceptual modelling is grounded on a foundational ontology that allows for more precise characterisations.

  Currently, ontology-driven conceptual modelling has not been applied to genomics. However, we will use ontology-based conceptual modelling to represent genomics and assess whether it is a better solution than traditional conceptual modelling. G2.5 led us to the following research question:

  RQ7 - How to conduct ontology-driven conceptual modelling in genomics?

  Ontology-driven conceptual modelling can be conducted in two ways: first, by creating a new conceptual schema from the very beginning; second, transforming an existing traditional schema into a ontology-based one. We call "ontological unpacking" the process of transforming a conceptual model without precise ontological background into an ontologically well-grounded one by using a foundational ontology that associates precise semantics to every concept of the initial model. To answer RQ7, we will apply an ontological unpacking to a portion of the CSG.

Finally, the last step will be to validate our contributions, which bring us to the last goal of this thesis:

■ **G3** - Confirm and validate our contributions and research results.

G3 led us to the following research questions:

RQ8 - To what extent are the contributions of this thesis useful in a human genomics context?
RQ9 - To what extent are the contributions of this thesis useful in an agri-food genomics context?
RQ10 - Does ontology-driven conceptual modelling capture domain particularities better than traditional conceptual modelling?

1.5 Thesis Summary

This thesis is structured in seven chapters, following the three main phases described by the research methodology (Problem Investigation, Treatment Design and Treatment Validation):

- Chapter 2 studies the evolution of genomics to understand its current data management problems and identify existing approaches to mitigate them.

- Chapter 3 presents the treatment design. Here, we describe the solution proposed to mitigate the problems presented in Chapter 2. Our work here includes updating and expanding the CSHG, the generation of the CSCG, the creation of the CSG, the creation of a method to create conceptual views from the Conceptual Schema of the Genome, and applying ontology-driven conceptual modelling in genomics.

- Chapter 4 validates the treatment presented in Chapter 3 in two real-world cases: precision medicine for humans and the improvement of crops using DNA modification techniques for agri-food. We have developed two conceptual model-based genome information systems, one per use case. This allowed us to validate the conceptual schema and the method since we created one conceptual view for each use case. The tools developed have been validated in various scenarios, including their use by students and domain experts. Additionally, after applying a process of ontological unpacking in our working context, we conducted two experiments to test if the particularities of genomics are better captured using ontology-driven conceptual modelling than traditional conceptual modelling.

- Chapter 5 reports conclusions and summarises the main contributions of this work regarding the scientific and academic communies. It also discusses future lines of research.

- Chapter 6 shows the impact of this Ph.D. thesis in terms of publications, teaching experience, participating of research projects, and organisation of congresses.
- Chapter 7 discusses future lines of research.

Part II

Main

Chapter 2

Problem Investigation

As Roel Wieringa states, the Problem Investigation phase allows researchers to identify which phenomena need to be improved and why [44]. The goal is to prepare for the Treatment Design phase by learning about the problem to be treated. That is, to study the main problems in genomics data management and the existing approaches to mitigate them (**G1**).

First, we must study the origin and evolution of genomics to understand its current data management problems. Then, we can identify the main data management problems (**RQ1**). Finally, the existing approaches used to deal with the identified problems can be reported (**RQ2**).

The chapter is divided as follows:

Section 2.1 – studies the evolution of genomics from its beginnings to the time of writing this thesis.

Section 2.2 – identifies the main challenges to be addressed in genomics.

Section 2.3 – describes how ontologies and conceptual modelling techniques deal with the challenges identified above.

Section 2.4 – reports the conclusions.

2.1 Evolution of Genomics

The History of Genomics

Genomics is a recent area when compared to other disciplines like architecture or mathematics. The history of genetics (see Fig. 2.1), which is the precursor of genomics, begins in 1869 with the Swiss physician Friederich Miescher [45], [46]. He was a student of Felix Hoppe-Seyler, one of the pioneers in a new discipline called "physiological chemistry" [47], who wanted to determine the chemical composition of cells. Through his experiments, Miescher showed that proteins and lipids were the main component of the cytoplasm and attempted to classify them. He also noticed a novel substance, DNA, which he called nuclein. He managed to isolate the DNA by developing new chemical protocols, reporting this discovery on February 26, 1869. Although Hoppe-Seyler was sceptical at first, he was convinced of his findings when he repeated the experiments and obtained the same results. Although neither Miescher nor his contemporaries could, at that time, fully grasp the significance of this discovery, he realised its great importance and future impact.

Fifty years later, Albrecht Kossel was awarded the Nobel Prize in Medicine for his contributions to cell chemistry, which included studying the composition of the cell nucleus and describing nucleic acids [48]. Kossel was convinced of the necessity of linking chemical constitution and biological function. He discovered the five nucleic acids, namely, adenine, cytosine, guanine, thymine, and uracil. Adenine was first isolated from the pancreas in 1885. Guanine was first isolated from a protein-free nucleic acid preparation in 1891. Cytosine and thymine were first isolated from "paranuclein" in 1893. Uracil was first isolated by Ascoli, one of his students, in 1901.

Thirty years after the identification of the nucleic acids that compose DNA, George Beadle and Edward Tatum published their most notable paper, titled "Genetic Control of Biochemical Reactions in Neurospora" [49]. They stated that the development and functioning of an organism is an integrated, interconnected system of chemical reactions, and they are controlled, in some manner, by genes. Their paper was motivated by the limitations of how their colleagues investigated genes (i.e., determining the physiological and biochemical bases of already known hereditary traits). By using an x-ray-based procedure in Neurospora[1], they demonstrated that the genes act by regulating definite chemical events (i.e., genes are responsible for synthesizing enzymes). As a result of their research, they won the Nobel Prize in Medicine in 1958.

[1] The Neurospora is a genus of fungus.

1871	The “nuclein” (later known as DNA) and proteins are identified in the cell nucleus.
1910	Nucleic acid bases are discovered.
1941	Gene are responsible for producing enzymes
1950	The pairing pattern of nucleic acids is discovered.
1953	The double helix structure of DNA is discovered
1955	RNA synthesis is described.
1957	The “central dogma” of biology is proposed.
1958	DNA replication is described.
1961	Protein synthesis is described.
1977	The Sanger sequencing technique is developed.
1986	The term *genomics* is coined.
1990	The Human Genome Project is launched.
1995	The first bacterium genome sequence is completed.
1996	The first first eukaryotic organism genome sequence is completed.
2000	The first plant genome sequence is completed.
2001	A first draft of the Human genome sequence is released.
2002	The first mammal genome sequence is completed.
2003	The Human Genome Project is completed.
2008	1,000 Genomes Project is launched.

Figure 2.1: A summary of the most remarkable milestones of genomics from its origins to the rise of Next-Generation Sequencing techniques.

Adenine, cytosine, guanine, and thymine were discovered in 1910, but how they interact with each other remained unknown until Erwin Chargaff discovered their pairing pattern [50], [51]. In Chargaff's words: *"it is senseless to formulate a hierarchy of cellular constituents and to single out certain compounds as more important than others. [...] It is impossible to write the history of the cell without considering the chronology of the cell. If this is done, nucleic acids will be found pretty much at the beginning"*. Chargaff was aware that DNA played a key role in the development of life by being the precursor and maintainer of cell functionality. He worked out the pairing pattern of the nucleic acids that make up DNA (i.e., he found that adenine pairs with thymine and cytosine pairs with guanine. Also, he found that DNA composition varies between species, but it does not vary between tissues of the same species.

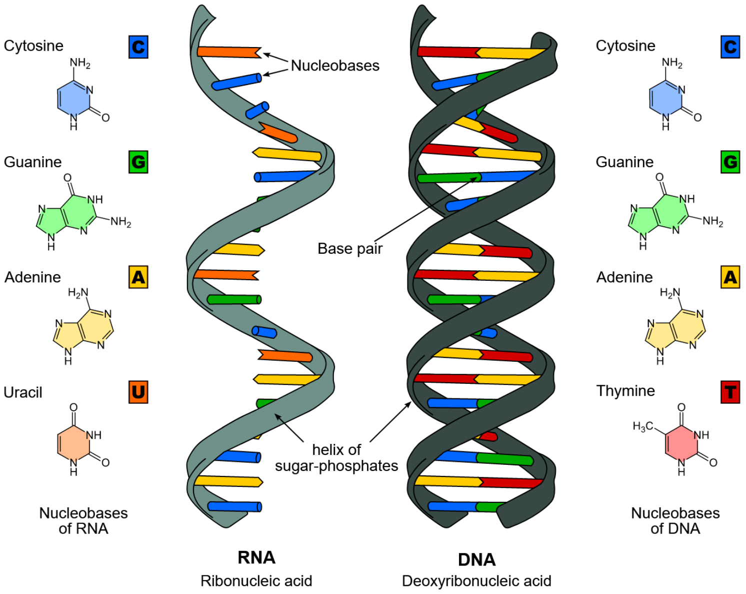

Figure 2.2: The chemical structure DNA and RNA nucleic acid bases.

The discovery of the pairing pattern of the nucleic acids comprised the most important single piece of evidence for the yet-to-be-described double-helical structure of DNA. Before the work of James Watson and Francis Crick, with contributions from Rosalind Franklin and Maurice Wilkins, the structure of DNA was proposed to be a three-chain structure. They proposed that DNA structure consists of two helical chains each coiled round the same axis [52] (see Fig. 2.2). In their words, the novelty of their proposal was: *"the manner in which the two chains are held together by the purine and pyrimidine bases"*.

Since the discovery of the structure of DNA by Watson and Crick, biochemists Arthur Kornberg and Severo Ochoa became interested in the nucleic acid synthesis mechanism. In one instance, in 1955, Spanish physicist Severo Ochoa isolated the enzyme polynucleotide phosphorylase (PNPase) from a bacteria, which plays a key role in RNA synthesis [53]. Ochoa was able to synthesise RNA *in vivo* using PNPase. While, in 1956, Arthur Kornberg isolated DNA polymerase I from a bacteria, crucial in the prokaryotic DNA replication process [54]. He synthesised complementary DNA chains using this enzyme.

In addition to the discovery of the double helix structure of DNA, Francis Crick also proposed the "central dogma" of biology in 1957 [55], [56]. This dogma, depicted in Fig. 2.3, explains how the genetic information flows (i.e., DNA makes DNA, DNA makes RNA, and RNA makes proteins). It also states that once information has passed into proteins, it cannot get out again. This means that transfer of information from protein to protein, or from protein to RNA is impossible.

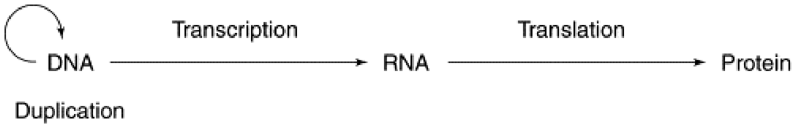

Figure 2.3: The central dogma of biology proposed by Francis Crick.

First, DNA can replicate itself when new cells are created. Second, DNA is transformed into a single-stranded RNA sequence in the transcription process. Third, the RNA is translated by the ribosome to create a polypeptide chain (i.e., a set of amino acids). The inner workings of these processes were unknown by the time the central dogma was proposed. Though, Matthew Meselson and Franklin Stahl described how DNA replicates one year later, in 1958. Through the Meselson-Stahl experiment, they discovered that DNA replicates in a semiconservative way [57]. Semiconservative means that two copies of the original DNA molecule are produced, and each of them contains one original strand and one newly synthesized strand (see Figure 2.4).

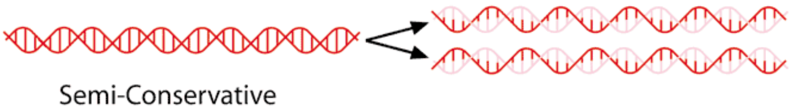

Figure 2.4: Semiconservative replication of the DNA.

The work of Crick, Ochoa, Kornberg and many others allowed Marshall Nirenberg, Har Gobind Khorana, and colleagues to crack "the code for life". Cracking the code for life means to comprehend how THE DREAM OF THE GENE BECOMES THE REALITY OF THE PROTEIN. Before their experiments, scientists knew that there are four amino acid bases (guanine, cytosine, adenine, and thymine) in DNA. Scientists also knew that there are twenty proteic amino acids. Consequently, Nirenberg and colleagues knew that the coding units could not be single (i.e., four combinations) or pairs (i.e., sixteen combinations). Through the Nirenberg and Matthaei experiment [58], they opened the door to answer a crucial question: *What code linked DNA sequences to each of the 20 amino acids that comprised proteins?* They concluded that the coding unit are triplets[2], and they deciphered the first three nucleic acids that translate for a specific amino acid: UUU codes for phenylalanine. This experiment caused a furious race to fully crack the genetic code and identify which triplets code for each amino acid and, as a result, the sixty-four codons were deciphered and linked to their corresponding amino acid by 1966.

Cracking "the code of life" laid the foundations for understanding how life works. Once we know (at a very basic basis) how DNA produces proteins, the next step is to start sequencing DNA at a more efficient rate, but the initial methods to sequence DNA were slow and complex. This limitation changed when Frederick Sanger developed the Sanger sequencing method [59]. The Sanger sequencing method used the DNA polymerase enzyme during the replication of DNA *in vitro* (see Figure 2.5 for a schematic example of the process). Soon, this method became the most widely used sequencing method. Frederick Sanger and his team used this method to sequence the first full genome sequence, the genome sequence of the ϕX174 virus[3]. As a result of his work, Frederick Sanger won the Nobel Prize in Chemistry in 1980.

The development of the Sanger sequencing method allowed for a more efficient and reliable sequencing technique and also was the catalyst for initiating ambitious projects like the Human Genome Project (HGP) [61], which was launched in 1990. The HGP aimed to sequence the 3 billion letters of the human genome sequence over fifteen years. It also aimed to identify and map every single human gene from a holistic perspective, including the structural and functional dimensions, among others. The first draft of the human genome sequence was released in 2001, and the project was completed in 2003, sequencing the complete human genome with 99.99 percent accuracy [62]. The HGP was the largest collaborative biological project and allowed scientists to obtain

[2] In the late 1950s, the term "codon" was coined to identify these triples. This term was popularised by Francis Crick later.

[3] This virus is a single-stranded DNA (ssDNA) virus that infects *Escherichia coli.*

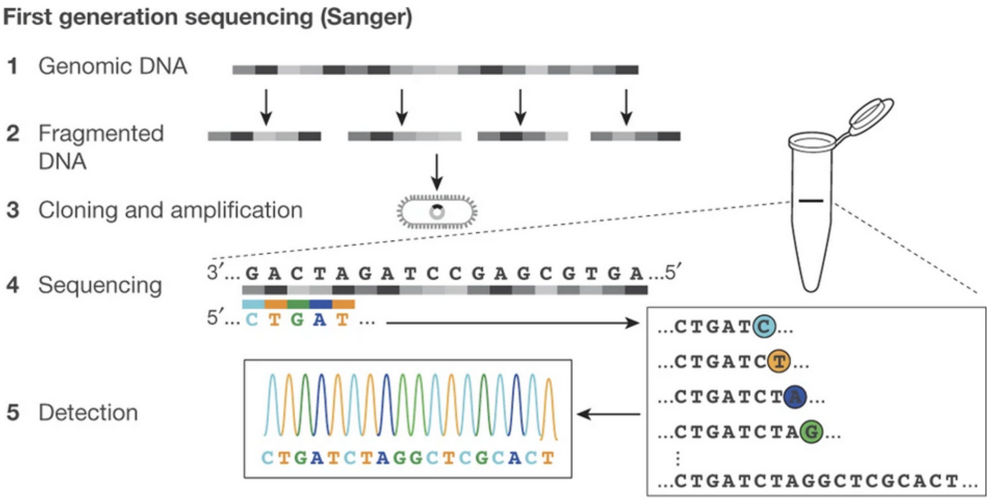

Figure 2.5: A schematic example of the Sanger process. Source: [60].

relevant information regarding our genome, from which we can mention the following [62]:

- The estimated number of human protein-coding genes is 26,383.
- Human protein-coding genes span about 27,000 bases on average.
- Human protein-coding genes contain, on average, 7.8 exons.
- The density of genes is greater in regions of high G+C than in regions of low G+C content.
- About 20% of the human genome is composed of gene-poor regions (regions of more than 500,000 bases without containing any gene). Though, this distribution is not uniform across the chromosomes.
- There is a strong correlation between CpG islands[4] and the first coding exon of genes.
- About 35% of the human genome is composed of repetitive DNA sequences.

[4] A CpG island is a DNA region of at least 200 bases where at least 50% of the sequence is a CpG site. A CpG site is a region of the DNA where a cytosine nucleotide is followed by a guanine nucleotide.

- About 2,909 regions have been identified as pseudogenes[5], although this number is likely an underestimate.

Multiple sequencing projects were carried out in parallel to the HGP, from which me can mention the following:

1995 The first complete sequence of the genome of a free-living organism (the bacterium *Haemophilus influenzae*) is published.

1996 The first complete sequence of the genome of a eukaryotic organism (*Saccharomyces cerevisiae*; i.e., the yeast) is published.

2000 The first complete sequence of the genome of a plant (Arabidopsis) is published.

2002 The first complete sequence of the genome of a mammal (*Mus musculus*; i.e., the mouse) is published.

2005 The first complete sequence of the genome of a primate (*Pan troglodytes*; i.e., the chimpanzee) is published.

2008 is considered a year of change and revolution regarding genomics because of two reasons. The first reason is that the 1,000 Genomes Project is launched. This project aimed to sequence the whole sequence of the genome of 2,500 people. The second reason is the significant decrease in sequencing costs due to the arrival of the Next-Generation Sequencing techniques. The advent of these techniques allows us differentiate between the *history* of genomics (i.e., a context of discovering the basics of genomics with limited capabilities of sequencing) and *current* genomics (i.e., a context in which the basics of genomics are well-understood and our sequencing capacity has grown exponentially).

Current Status of Genomics

We can consider two dimensions regarding the scientific knowledge of the *current* status of genomics: the **biological dimension** (i.e., our understanding of the genome) and the **technological dimension** (i.e., advances associated with sequencing capacity).

Considering the **biological dimension**, the most remarkable event was the fall of the central dogma of biology. This famous dogma is an over-simplification

[5] A pseudogene is a non-functional copy that is very similar to a normal gene.

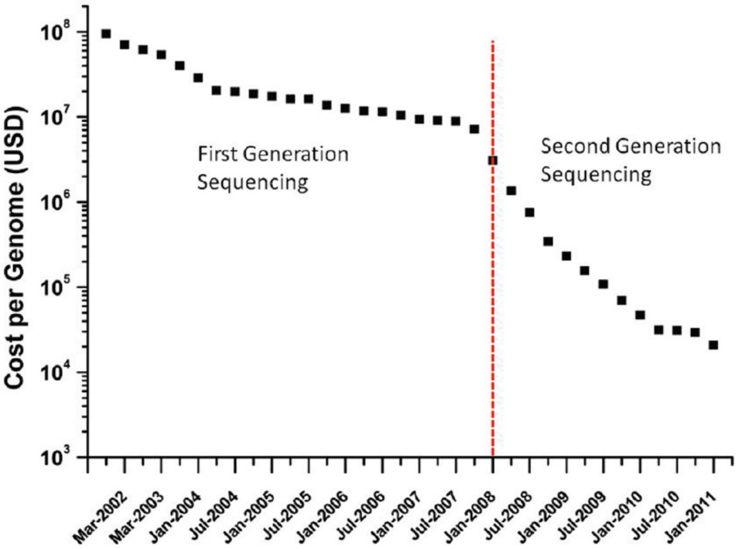

Figure 2.6: Approximate start date of Second Generation Sequencing depicted on a graph that shows the estimated cost of sequencing a Human Genome. Source: [63].

of the complex network of processes that drive life. For instance, the central dogma claims that genetic information flows through macromolecules (i.e., from DNA to RNA to proteins). However, macromolecules alone are not sufficient to sustain life [17].

Since the elaboration of the central dogma, our comprehension of cell function, biological chemistry, and many other fields has grown exponentially as radical discoveries arose [64]. Such discoveries violate the central dogma at multiple points, posing the need for revisiting it. Many unexpected and even “forbidden” activities have been discovered:

- Reverse transcription: Because of the reverse transcriptase process, information can flow from RNA to DNA [65].
- Post-transcriptional RNA processing: The information contained in RNA has many potential inputs apart from the original DNA template [66]. RNA is modified after being transcribed from DNA by several processes like splicing, cleavage, etc [67]. Even more, trans-splicing can join two different RNA sequences [68].

- Catalytic RNA: RNA molecules can undergo structural changes, which means that they have catalytic processes analogous, in many ways, to proteins [64]. Consequently, RNA molecules have a more critical role in determining cellular characteristics, rather than being limited to being a template for protein-coding processes.
- Genome-wide transcription: The original dogma discriminated between protein-coding DNA and non-protein-coding DNA (also called "junk" DNA) that was assumed to have no purpose. However, virtually all DNA is transcribed and provides specific functionality [69].
- Post-translational protein modification: Like with RNA, proteins can be modified after translation, altering their functionality [64].
- DNA repair: The original dogma assumed that DNA replication was inherent to DNA and its own machinery, but several DNA replication mechanisms exist at the protein-level [70].

In addition, cells acquire information about their environment, keeping track of several internal processes [64]. Such information allows cells to modify their internal processes (including replication, transcription, or translation). This goes against the initial vision of DNA as an isolated, unidirectional process that remained unaltered. Another aspect of the original dogma is that it considered genes and proteins as unique, unalterable, and unitary entities. This unitary perspective has shifted because large amounts of composite components exist in our bodies, such as protein complexes [64].

Finally, the central dogma assumed that DNA is a stable structure that rarely changes. However, DNA changes almost constantly due to epigenetic factors or errors in replication processes [71], and numerous systems dedicated to DNA restructuring and repair exist [72].

All of these facts together mean that the central dogma needs to be questioned and, therefore, reformulated. Attempts to do so exist, like the one presented in [73] (See Fig. 2.7). The reader should bear in mind that there is currently no agreement regarding the exact processes that should include such a reformulated dogma, but we think it is interesting to show a proposal to illustrate the complexity of life and its inner working mechanisms.

Considering the **technological dimension**, sequencing costs as well as the time required to do the sequencing have decreased significantly [74], [75]. Consequently, the amount of genomics data that is available has increased substantially [76]. Here, we report the advances in sequencing techniques and the

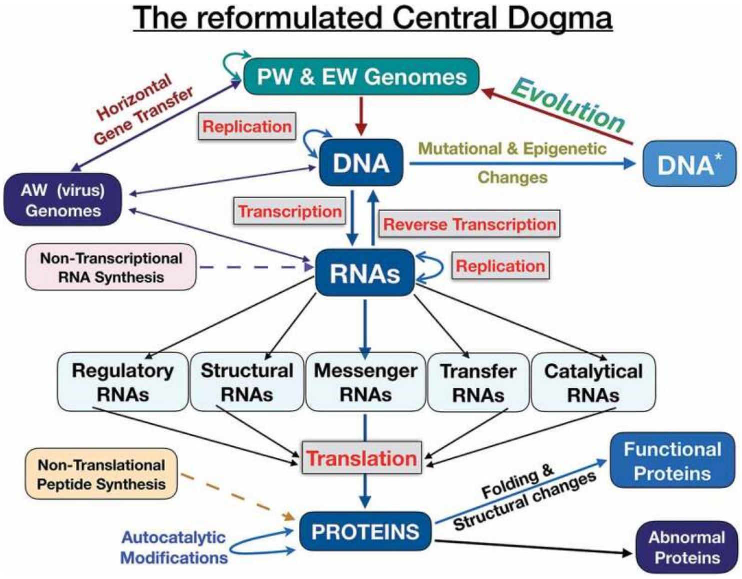

Figure 2.7: A reformulated central dogma of biology. Source: [73].

resulting amount of generated genomics data; the significance of this is discussed in depth in Section 2.2. Sequencing the human genome cost nearly 100 million dollars in 2001; today, this has been reduced to one hundred dollars (see Figure 2.8) [77]. The costs reported in Figure 2.8 refer to the generation of a high-quality draft of a whole genome sequence. There are other commercial options at even lower costs, albeit with slightly decreased quality.

The technologies used to perform DNA sequencing can be divided into three generations. Although there are discussions regarding the exact dates that delimit these generations, there is a broad consensus in the scientific community regarding the time periods: [60], [78]:

(1977 - 2008) First Generation — Sequencing of DNA through the use of cloning vectors.

(2008 - present) Second Generation — Increased throughput by parallelising many reactions.

(2010 - present) Third Generation — Direct sequencing of single DNA molecules, avoiding the need for DNA amplification.

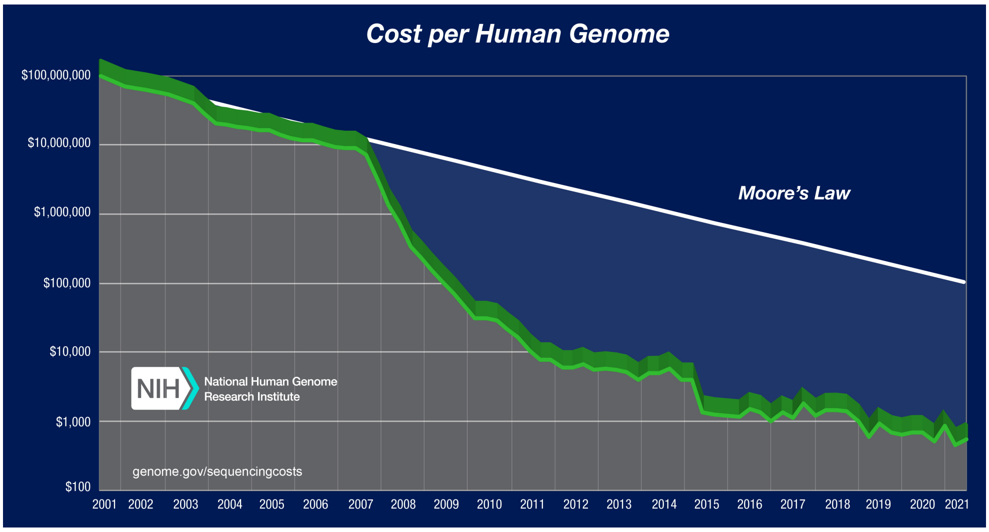

Figure 2.8: The Cost of Sequencing a Human Genome. Source: National Human Genome Institute.

The focus of this thesis is not on sequencing techniques. However, we consider that providing the reader with a brief introduction to the evolution of sequencing techniques helps to provide and understanding of the current issues when managing genomics data and provides a rich context to the problems and the solutions that have been developed.

As a reminder, since First Generation DNA Sequencing Techniques (FGT) require the direct action of DNA polymerase to produce the observable output, they are considered sequence-by-synthesis techniques. First Generation DNA Sequencing Techniques inferred nucleotide identity by using radio or fluorescently labeled modified nucleotides before visualising them with electrophoresis. There are two main techniques in this first generation, namely, Sanger sequencing and Maxam-Gilbert sequencing [78], [79]. Sanger Sequencing consists of using one strand of the double-stranded DNA as a template to be sequenced, which is made using chemically modified nucleotides. Maxam-Gilbert is another sequencing method that was known as the chemical degradation method. It is based on the cleaving of nucleotides by chemicals instead of DNA cloning. It is noteworthy to remark that the Maxam-Gilbert method is considered dangerous because it used toxic and radioactive chemicals. It is also important to remark that FGT are still used but is limited to specific activities like small-scale experiments or sequencing regions that can not be easily sequenced using more recent techniques (e.g., highly repetitive DNA) [80].

A paradigm shift underlies Second-Generation DNA Sequencing Techniques (SGT) [78], [81]. They differ from FGT in several ways [79]: SGT use multiplexing, whichs allows for a significant increas in throughput by parallelising many reactions, and it also sequences DNA using light detection methods that measure the light released when a nucleotide is synthesised (this method is called pyrophosphate synthesis because pyrophosphate is released when nucleotides are synthesised). In addition to these differences, SGT are considered to be sequence-by-synthesis techniques like FGT.

In general, SGT can be divided into three steps [60], [82]. The first step is to prepare the template where the sequencing is to be performed. The second step is to do considerable amounts of parallel amplification. The third step is to perform sequencing and alignment. In the first step, a complex library of DNA templates is densely immobilised onto a two-dimensional surface instead of one tube per reaction. In the second step, DNA molecules are clonally amplified in an emulsion, producing clusters of clonal DNA populations. In the third step, the clonally amplified sequences are read in a highly parallelised manner. Figure 2.9 shows a schematic example the second and third steps.

With SGT, many millions of short reads are generated in parallel, it is significantly faster and cheaper when compared to FGT, and the sequencing output is directly detected without the need for electrophoresis [63]. However, finding out the number of the same nucleotides that are in a row at a given position might be difficult because noise can alter the intensity of the light that is released [63]. There are issues related to the biases introduced by cloning amplification and dephasing[6] [84].

Other than the drawbacks of using SGT mentioned above, there are additional sequencing applications that are relevant but that are beyond the reach of SGT [84]. However, Third-Generation DNA Sequencing Techniques (TGT) overcome the issues associated with using SGT (e.g., biases introduced in clonally amplification). These techniques (i.e., TGT) do not require clonally amplification, which allows for a faster and more efficient DNA sequencing [79]. Apart from negating the requirement for DNA amplification, the main characteristics of TGT are that they allow for the direct sequencing of single DNA molecules and real-time sequencing [78].

[6] The reactions used to clonally amplify DNA gradually lose its synchronization within the molecular colony, causing sequencing errors [83].

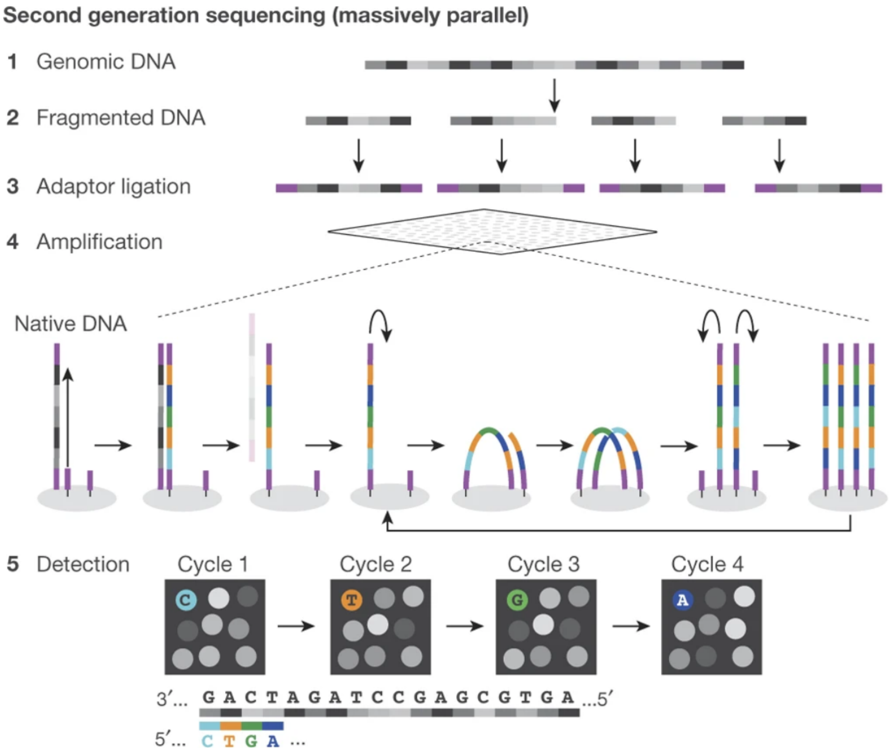

Figure 2.9: A schematic example of a second-generation DNA sequencing technique. Source: [60].

Currently, the two most promising TGT are Single Molecule Real Time sequencing (SMRT sequencing) and nanopore sequencing [63], [79], [84]. SMRT relies on sequencing a single molecule in real time by synthesis methods. Nanopore sequencing measures translocation of nucleotides cleaved from a DNA molecule across a pore. This measurement is driven by the force of differential ion concentrations across the membrane of the pore.

The use of SGT and TGT represent a significant challenge regarding data generation [79]. Genomics has become a subfield of big data science that has exceeded other big data domains like astronomy or social media [85], [86], posing unprecedented challenges in data acquisition, data storage, data distribution, and data analysis [86]. For example, different estimations predict that between 2 and 40 exabytes of genomics data will be generated within the next decade [87]. In this situation, our ability to generate data has outpaced our ability to decipher the information it contains.

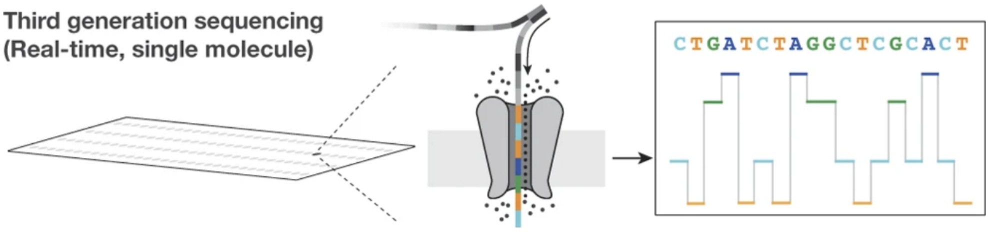

Figure 2.10: A schematic example of a third-generation DNA sequencing technique. Source: [60].

This is the main point that justifies our research efforts and the importance of applying conceptual modelling to the challenge of deciphering the language of life. As more and more data are available, we must understand, interpret, and manage such data correctly. The study of this challenging and complex "genomics data science" problem is our main purpose.

Although many challenges need to be addressed, being able to generate such large amounts of data has already resulted in significant great benefits. For example, in oncology, the impact of having immense amounts of data has been transformational [88]. More concretely, our understanding of how DNA modifications can lead to uncontrolled cell division has been expanded greatly. Sequencing, integrating, and abstracting these data allow researchers to determine what pathways drive cancer when they are dysregulated and how.

Extending our understanding of these changes as they impact cancer biology has been enabled by sequencing various classes of RNA in the cancer cell, the integration of these data with identified DNA alterations, and abstraction of this information to the cellular pathways that drive cancer when they are dysregulated. In line with that, Ana-Teresa Maia et. al. state that such an amount of genomics data related to cancer will allow for identifying genetic predisposition changes, prognostic signatures, and cancer driver genes [89].

However, despite these examples, several issues regarding genomics data still need to be addressed: precise conceptual characterisation of genomics data, generation of sound and ontologically well-grounded genomics-based knowledge, more sophisticated analysis tools, higher processing capacity, data integration, data harmonisation, etc. These challenges are often referred to as the *genomic data chaos*. The next Section provides a thorough analysis of what the genomic data chaos is, and how it is characterised.

2.2 Genomics Data Management Status

The increase in the efficiency of NGS techniques and their reduction in costs has given rise to a context where genomics has become a big data and data-intensive domain problem, whereby correct data management is an open challenge. The main issues to be considered for managing genomics data are the following:

Volume – The first issue is related to the existing large amount of genomics data and the fact that the rate at which it is being generated is increasing drastically. At this moment, the cost of sequencing a whole genome human has broken the $1,000 barrier, ranging between $300 and $600 . Besides, sequecing a wholse genoe human sequence takes less than a week. Considering that a single human genome sequence takes up to 100 GB of storage, sequencing every living human (an estimated amount of eight billion people) would require almost one zettabyte of storage (that is 10^{21} bytes). To put it in context, 64 zettabytes of data were created, captured, copied, and consumed worldwide in 2020, according to the International Data Corporation (IDC). Even more, single-cell sequencing techniques aim to obtain the DNA sequence of a single cell [90]. Humans are composed of approximately 30 trillion cells approximately [91], and cells die and are born over time, potentially appearing with new DNA variants in their sequences. Therefore, it is not difficult to imagine the amount of space that would be required to sequence and store even a small portion of the cells of every human being on a regular basis.

Heterogeneity – The second issue is related to the existing amount of data sources. Technological advances expanded the available genomics data to a great extent, from whole-genome DNA sequences to epigenetic-related data. Such genomics knowledge is spread across over hundreds of heterogeneous databases (1,637 databases containing genomics data existed in 2020 according to [92]). These databases have different sizes, formats, and structures. More importantly, they focus on different genome dimensions, making integrating heterogeneous data a highly complex problem (also documented in the next item).

Lack of interconnection – The third issue is related to the isolation of the existing genomics data. On one hand, this issue is related to the previous challenge. There are thousands of databases isolated from each other, which complicates data integration. While, on the other hand, new high-throughput omics and mass spectrometry allowed for generating large amounts of epigenomics, transcriptomics, proteins, and metabolomics data (among many others) [93], [94]. The existence of so numerous data types with a wide variety of notations and formats makes their effective combination a near-impossible task. Thus, data integration is a fundamental yet not-solved activity [95].

Evolution – The fourth issue is associated with the pace at which new genomics knowledge is generated. Our knowledge of genomics, along with the rest of omics, is constantly evolving. There is a high variability dimension associated with genomics data. To illustrate, let us focus on a particular area of genomics: variation classification. In the clinical domain, it is common to associate variations and diseases with a given degree of confidence to improve patients' diagnosis and treatment, but variant reclassification over time is not a rare event [96]. For instance, 70% of rare genetic variations associated with the inherited arrhythmogenic syndrome have been reclassified [97]. Another example can be found in genes, where, in 2021, the American College of Medical Genetics and Genomics (ACMG) expanded the list of genes, by an additional 13, when conducting genomics studies 13 additional genes in 2021 [98].

This situation is known as *the genomic data chaos*, a concise term that shows the existing problems, their relevance, and their multiple dimension.

2.3 Genomics Data Management Strategies

There are many strategies aimed at improving genomic data management, but we only consider those based on conceptualisation. With that in mind, the most relevant approaches to dealing with the problems arising from the genomic data chaos are domain ontologies and conceptual modelling, which are different but complementary. Both perspectives emphasise the importance of making explicit conceptualisation a common practice, which ensures better understanding and communication.

Domain Ontologies

Domain ontologies define abstract conceptualisations of the essential knowledge associated with a specific domain in an extendable and practical way. Every domain ontology is built upon a foundational ontology, which provides a high-level characterisation about the most fundamental concepts. The Basic Formal Ontology (BFO) [99] or the Unified Foundational Ontology (UFO) [100] are examples of foundational ontologies.

These “genomics ontologies” provide a shared thesaurus of terms that are ordered hierarchically. They make up a broad family of solutions that are currently the most popular ones. The Ontology Lookup Service[7] retrieves 273 ontologies comprising 7,362,786 terms. One significant representative of this approach is the Open Biological and Biomedical Ontology (OBO) Foundry [101]. The OBO is an entity whose mission is to provide a set of design ontology principles. The OBO ontologies are *loosely hierarchical directed acyclic graphs* (e.g., a concept may have more than one parent term). They organize domain knowledge into two dimensions: granularity and relation to time.

However, a number of limitations arise when using genomics ontologies. The first limitation is the discontinuation of ontologies; hundreds of so-called ontologies have been defined following their principles, but dozens are already obsolete. Another limitation is the fact that they only cover a part of genomics, which complicates data integration and hinders the whole picture. For instance, there are ontologies to characterize phenotypes (HPO), gene functionality (GO), genome sequences (SO), proteins (PRO), and variations (VO). These five ontologies reduce domain heterogeneity by providing well-defined standards for specific domain concepts. However, they do not provide a standard, clear definition of the concepts that they explore [102].

In addition, there is not an explicit link among these ontologies, meaning that different ontologies can characterise two related concepts without specifying how they are linked, or one common concept can be represented differently in alternative ontologies.As an illustration : phenotypes are characterized in multiple ontologies such as the Unified Phenotype Ontology (UPO), the Mammalian Phenotype Ontology (MP), or the Mouse pathology Ontology (MPATH). The term *apoptosis* is present in the three ontologies, but they are not linked. Even more, the Human Phenotype Ontology (HPO) and the Neuro Behaviour Ontology also describe phenotypes, complicating integration processes even more.

Managing a particular concept of genomics that is described by one ontology (a “vertical” dimension) works reasonably well. Nevertheless, problems arise when establishing semantic connections among different concepts described by multiple ontologies (a “horizontal” dimension). For instance, knowing why a specific change in the genome produces the clinical manifestations of a disease requires navigation through the different concepts:

[7] `https://www.ebi.ac.uk/ols/index`

- The chromosomal elements affected by the variant (a gene, transcripts, a protein, its isoforms, etc.).
- The functions that these elements perform (transcription regulation, ion transport, protein degradation, etc.).
- The biological processes and reactions where these functions are involved (immunological response, tissue growth, etc.).
- The consequences of the malfunction of these processes (recurrent infections, growth retardation, etc.).

Apart from what we have mentioned above, there are additional pitfalls that are ontology-specific such as missing relevant terms, typos, or incomplete term definitions of missing constraints [103]. Finally, domain ontologies have scalability issues because the more an ontology grows, the harder is to maintain and avoid including duplicated terms.

In conclusion, domain ontologies are well-established solutions that are broadly used and considered to be useful approach that helps to reduce domain heterogeneity and improve data integration and information exchange. However, they present a significant set of limitations (discussed above) that require different approaches to mitigate them.

Conceptual Modelling

Conceptual modelling focuses on understanding and representing cognitions about the world. Computer science uses this mainly for the development of information systems for a range of applications. The use of conceptual models by means of a given formalism (e.g., an Entity-Relationship, Object-Role Modeling or UML class diagrams) is a powerful tool for understanding and communicating complex domains regardless of the research area. It clearly identifies the relevant entities involved and the relationships among them. [104]. With such a holistic perspective, conceptual modelling applied to the genomic data domain can help to solve the limitations associated with the use of domain ontologies.

Conceptual modelling has been explored by some authors in the past:

- Chen et. al. proposed an Object-Protocol Model [105] to bring an example of a combination of protocol and object constructions in a framework for the genomic domain that allowed modelling of objects and experiments (protocol).

- The DNA Databank of Japan (the DNA Data Bank of Japan) designed and developed a new version of their Nucleotide Sequence Databank using conceptual modelling to facilitate the acquisition and maintenance of genomics data [106].
- The work presented in [107] introduced a cooperative computing environment for the analysis and annotation of DNA sequences. This work was developed by applying an object-based model.
- Paton et. al. [108] described the genome from different perspectives using a conceptual modeling-based approach. Their work included the description of the eukaryotic cell genome, the interaction between proteins, the transcriptome, and other genetic components.
- In [109], the principles of Conceptual Modelling (CM) were applied to describe particularities of protein structures in their 3D form.

These approaches still focus on specific parts of the domain that are not connected to each other, and they do not provide the required global view to understand complex biological systems. However, more recent attempts to provide a sound, CM-based solution from a more holistic perspective are being developed:

1. Bernasconi et al. developed the Genomic Conceptual Model (GCM) [110], an entity-relationship diagram designed to integrate genomic signals (e.g., DNA mutations, the expression of gene activity, or DNA's structural rearrangement, among others) represented in several heterogeneous data formats. This work improved data sharing by standardising metadata across different data sources. Based on this conceptual model, a novel architecture for large-scale genomic metadata integration called METABASE was created [111]. Additionally, the GCM served as inspiration for creating an additional conceptual schema, called the Viral Conceptual Model (VCM) [112], for expressing the characteristics of viral sequences.
2. The Global Alliance for Genomics and Health (GA4GH) [113] is an alliance with the mission to improve data sharing of clinical and genomics data by means of conceptual modelling. The main standards developed by this alliance are "Phenopackets" [114] and the "Variation Representation Specification" [115]. The Phenopackets standard provides an appropriate way to communicate bioinformation for the purposes of research, diag-

nosis, and treatment. While the Variation Representation Specification standard is used to facilitate and improve the representation and sharing of variations.

All of the above mentioned works are conceptual modelling efforts that have succeeded in achieving their goals. However, genomics is a particularly complex domain, and it is very difficult to capture and represent its particularities appropriately. Because of the genomics data management issues, there are several concepts that have intentionally abstract definitions or have multiple underlying interpretations.

In other complex domains, recent approaches tried to capture domain particularities by building their modelling efforts upon a solid ontological foundation (i.e., a foundational ontology). Several foundational ontologies exist, such as Basic Fomal Ontology (BFO) [99], Suggested Upper Merged Ontology (SUMO) [116], Business Objects Reference Ontology (BORO) [117], Descriptive Ontology for Liguistic and Congnitive Engineering (DOLCE) [118], General Formal Ontology (GFO) [119], or Unified Foundational Ontology (UFO) [100].

Using ontology-driven conceptual modelling is considered to be more advantageous than using traditional conceptual modelling [120]. Some studies have considered the differences between traditional and ontology-driven conceptual modelling in various domains [121], [122] and from a theoretical perspective [123].

2.4 Conclusions

Throughout this chapter, we tackled **G2**. First, we have studied the main problems associated with genomics data management, answering **RQ1**: Which problems arise when working with genomics data? After three generations of sequencing machines, the throughput and speed of genome sequencing have grown exponentially, allowing for unprecedented amounts of genomics data. These data have been generated following several guidelines and standards and are spread over hundreds of data sources. Additionally, there continues to be breakthroughs and discoveries further expanding the field of genomics. As a result, genomics data management and knowledge extraction have become inefficient and ineffective.

From a conceptualisation perspective, two approaches have prevailed to address this situation: domain ontologies and conceptual modeling. Domain ontologies define hierarchical graphs of related terms and concepts and are the most popular choice. Ontologies allow the most relevant terms of a particular dimension of genomics to be defined (i.e., they provide a vertical-oriented solution).

Conceptual modelling, the second approach, explicitly describes the most relevant terms in a domain and their relationships. Conceptual modelling has become increasingly important over time and it can complement domain ontologies by describing the domain from a more holistic perspective. It can also connect different dimensions of genomics. (i.e., they provide a horizontal-oriented solution). There are two approaches to conceptual modelling: traditional conceptual modelling and ontology-driven conceptual modelling. The latter has extended modelling capabilities because of the types and definitions provided by its underlying foundational ontology. The study of these two approaches allowed us to answer to **RQ2**: What existing approaches can be used to mitigate the identified problems?

We have thoroughly studied the main problems in genomics data management and how they are mitigated. This thesis applies conceptual modelling techniques to better share and represent knowledge, improve genomics data management, and generate knowledge more efficiently. In the next section, we elaborate on how we used conceptual modelling to achieve our goals.

Chapter 3

Treatment Design

In the previous chapter, we explored the details and particularities of the context of this research, studied the main problems that can arise when managing genomics data, and identified the main approaches to mitigate them. In this chapter, we describe the main contributions of this thesis, which aims to generate conceptual modelling artifacts to improve genomics data management. (**G2**). This high-order goal is divided into five subgoals.

First, we extend the current version of the CSHG to be used in real-world use cases (**G2.1**). Second, we explore the conceptualisation of the genome for another species rather than humans, creating the CSCG (**G2.2**). Third, based on the experience accumulated from the CSHG and the CSCG, we generate a conceptual schema that is species-independent called the CSG (**G2.3**). Fourth, we provide a method to facilitate the adoption of the CSG to work with genomics use cases (**G2.4**). Fifth, we identify another artifact for better representing genomics, generating a schema that is based on a foundational ontology (**G2.5**).

The chapter is structured as follows:

Section 3.1 – extends the Conceptual Schema of the Human Genome.

Section 3.2 – creates the Conceptual Schema of the Citrus Genome.

Section 3.3 – creates the Conceptual Schema of the Genome.

Section 3.4 – creates the ISGE method.

Section 3.5 – performs the ontological unpacking.

Section 3.6 – reports conclusions.

3.1 The Conceptual Schema of the Human Genome

The treatment design phase starts by tackling **G2.1**. The last version of the CSHG (i.e., Version 2) was presented in the thesis of Ph.D. José Fabián Reyes Román [124]. Version 2 of the CSHG has been used in different lab contexts for several years, allowing the gathering of valuable feedback.

Here, we analyse this feedback and consider the most relevant scientific discoveries in genomics since the development of Version 2 and use this to identify the five dimensions that needed to be improved upon, the result of which led to Version 3 of the CSHG.

3.1.1 Introduction

To generate Version 3 of the CSHG, five dimensions needed to be improved:

1. **Improving the independence of the CM from technological implementations**: There were concepts tied to specific solutions or technological implementations. For instance, the GENE class had the *id_ hugo* attribute, which corresponded to an identifier provided by the HGNC database[1]. This identifier is not universal and is not shared among other data sources; thus, limiting data integration and flexibility for including additional information in the CSHG. *Should specific data source attributes be used in our CHSG, or should we use a more agnostic approach?*

[1] https://genenames.org

2. **Considering multiple assemblies**: Version 2 of the CSHG did not allow for genome versions of the sequence of reference to be modelled. This conceptualisation did not consider representing more than one reference sequence coexisting in time, which is the current situation due to the technological limitations of sequencing technologies. A relevant consequence of having multiple reference sequences is that the position of a variation is not unique but relative to the reference sequence being considered. *How to model multiple assemblies (coming from different reference sequences) in the schema?*

3. **A new way of representing variations**: The CSHG represented variations with respect to their type and frequency among populations, which exhibited a number of limitations, modelled the information redundantly, and was too complicated. Regarding the limitations, a variation could be either a polymorphism or a mutation in terms of frequency. However, only a specific type of polymorphism (i.e., the SNP) was associated with genotype and population information. This means that genotype and population information could not be associated with a variation that was not an SNP. With redundancy, variations were represented at least twice (i.e., by frequency and population). Even worse, a variation could be represented more than twice if its frequency varies among populations. In terms of complexity, it was necessary to define exclusive disjunction XOR rules to ensure data correctness. For instance, a variation that was not precise (type) did not have to be represented as a polymorphism (frequency). *Can the representation of DNA variations be simplified?.*

4. **Improving the representation of the effects caused by variations**: The CSHG only represented the phenotypic effects caused by variations, but not lower-level effects in the organism, such as alterations in the structure of a protein. Representing these effects would increase the completeness of our model by providing an additional means of describing the consequences of DNA variations in our body. *What are the consequences of DNA variations at a low level?.*

5. **Extending the representation of gene products**: The CSHG missed some relevant concepts and relationships associated with the transcription process. For instance, messenger RNA, which is defined as the intermediate product between the genome and the proteome, was not represented explicitly. In addition, the model only considered the coding protein transcription process, which is a correct but incomplete assumption because non-coding RNA is also obtained through transcription. *How to represent the outcomes of transcription correctly?.*

The next section describes each of the problems listed above in more detail, it explains how they have been solved, and it discusses the associated ontological commitments. The development of these five ideas led us to the next version of the CSHG (i.e., Version 3), which is the first original contribution of this thesis.

3.1.2 Improving the independence of the CM from technological implementations

The CSHG lacked flexibility because some of the classes had attributes associated with specific solutions. This situation predisposed domain users to use only those data sources whose identifiers were represented in the CS. However, depending on the working context, some of these attributes might not be used or could not be obtained. Besides, working with new data sources represented a problem because the CS and its database implementations had to be updated in order to use the new sources. The CSHG needed to avoid referring to specific solutions because domain users focused on how to deal with data management issues rather than generating knowledge. The classes tied to specific data sources were ChromosomeElement, Variation, and Population ones.

The Variation class is an excellent example that shows the problems of the approach used in Version 2 of the CSHG. Variations were linked to one data source (represented in the CSHG with the DataBank class) and had four attributes representing technological-dependent identifiers, namely, *rs_identifier*, *nc_identifier*, *ng_identifier*, and *id_hugo*. In addition, the *other_identifiers* attribute contained additional identifiers. With this approach, we could not determine how many identifiers a variation had, and there was a loss of information because the identifier was not associated with its corresponding data source.

We implemented an abstraction mechanism that allows for representing any data source-specific identifier. The ChromosomeElement, Variation, and Population classes are no longer directly linked to a single data bank. Instead, a new concept for describing technological-dependent identifiers links these classes to the corresponding data bank (See Fig. 3.1). After the update, these classes no longer require attributes that are associated with a specific solution, it is easy to determine how many identifiers they have, and there is no loss of information since each identifier is linked to its corresponding data source.

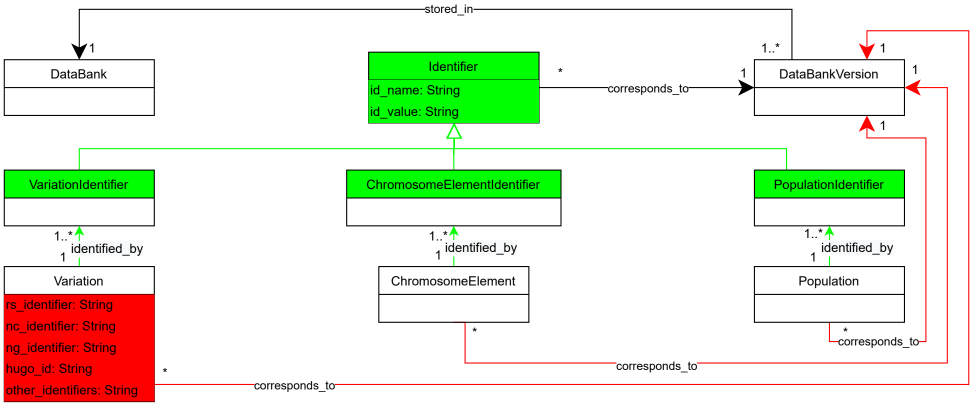

Figure 3.1: Changes for providing a more agnostic approach to technological-specific identifiers of genomics concepts. Additions are depicted in green while deletions are depicted in red.

This approach eliminates any possible bias as it does not explicitly represent attributes associated with specific solutions. It also provides domain users with the ability to represent any new data source without having to update the CS. We are now able to represent any technological-dependent identifier.

3.1.3 Considering multiple assemblies

The second version of the CSHG modelled the human genome assuming that there is only one sequence of reference (i.e., each chromosome has a unique sequence). In this approach, regions in the DNA, such as chromosome elements or variations, were located in only one position in the reference sequence. However, due to technological limitations in the sequencing process, multiple versions of the sequence of reference of the human genome coexist in time. It is common practice to work with more than one simultaneously. For instance, the rs11571636 variation [125] has different positions depending on the version of the human genome sequence being considered: chromosome 13, position 32,905,026 in the GRCh37[2] assembly; and chromosome 13, position 32,330,889 in the GRCh38[3] assembly.

[2] Reference sequence of the human genome, build 37 `https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/`

[3] Reference sequence of the human genome, build 38 `https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/`

These technological limitations in the sequencing processes mean that it is impossible to read the entire genome sequence at one time. Instead, the DNA is split into several parts that are read multiple times, obtaining a set of overlapping sequences. The sequences with the best reads are joined to create contigs, which, in turn, are concatenated to form scaffolds (i.e., non-contiguous contigs that might be separated by gaps of known length but unknown sequence) [126]. Finally, all of these scaffolds are grouped into what is called an assembly, which contains the chromosome sequences, like the GRCh37 or the GRCh38 [127].

We considered three different proposals to update the CSHG. In the first approach, a new instance of the CS is generated for each assembly. Although this approach would allow us to work with multiple assemblies, it was discarded for three reasons. First, the concept of the assembly would not be explicitly represented in the CS. Since it is a relevant and widely used concept in genomics, it should be described. Second, domain users work with multiple assemblies simultaneously because variations of interest can be identified in one or several assemblies. This fact should be appropriately represented in the CS. Third, this approach would introduce unnecessary redundancy since the rest of the data is duplicated on any new instance of the CS.

In the second approach, the assembly is included in the CS, instantiating a set of chromosomes per assembly. This approach would solve the three problems of the first approach: **i)** the assembly would be represented in the CS, **ii)** multiple assemblies could be represented simultaneously, and **iii)** unchanged parts of the CS would not be duplicated. However, this approach has a conceptual weakness: the notion of the chromosome is unique and should not be instantiated multiple times. What changes on each assembly is the sequence of the chromosomes, not the chromosomes themselves.

The *sequence* attribute is extracted from the Chromosome class in the third approach, which is the one we adopted (see Fig. 3.2). This approach solves the conceptual weakness of the second approach while keeping its advantages since chromosomes are instantiated only once. After the update, each chromosome may have many reference sequences associated with the corresponding assembly while modeled as a unique entity. This approach is the most suitable and conceptually accurate regarding assemblies, chromosomes, and their associated sequences.

Having chromosomes with multiple sequences of reference means that the ChromosomeElement (i.e., regions of the chromosome reference sequences with specific functionality or biological interest) and Precise (i.e., variations whose position is precisely identified) classes are no longer located in one and only one position. We needed to represent this fact in our schema; otherwise, contextual information would be lost because it is impossible to know the locations of these classes among the different assemblies. We decided to extract the locations of these classes into a new class (i.e., the VariationPosition and ChromosomeElementPosition classes). After the update, the positions of these classes among the assemblies are adequately represented (i.e., they are no longer located once, but rather once per assembly).

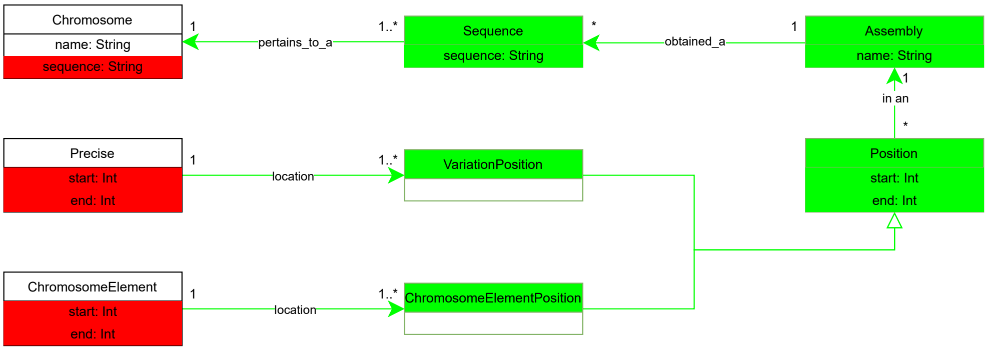

Figure 3.2: Changes for representing the concept of assembly adequately. Additions are depicted in green while deletions are depicted in red.

Now, the CSHG can represent multiple genome assemblies with their corresponding reference chromosome sequences and identify the positions of chromosome elements and variations in multiple assemblies. Since sequencing technologies are improving rapidly, our schema will be able to deal with the increasing number of assemblies.

3.1.4 A new way of representing variations

The CSHG described variations based on its frequency and type. Regarding the classification based on frequency, a variation was considered a Polymorphism if it appears with a higher than one percent frequency in a given population; otherwise, it was considered a Mutation variation. Polymorphisms could be a Single Nucleotide Polymorphism (SNP) if there is a one nucleotide change; otherwise, they were a copy number variation (CNV).

Regarding the classification based on type, a variation was considered Precise if its position in the reference sequence was precisely identified; otherwise, it was considered Imprecise. A precise variation could be specialised into an Insertion (i.e., a novel region is inserted in the genome), a Deletion (i.e., an existing region is removed from the genome), an Indel (i.e., an insertion and a deletion), and an Inversion (i.e., a region of the genome that is inverted).

Four issues were identified as a result of how variations were classified in Version 2 of the CSHG:

1. The Population (i.e., a set of individuals that share a characteristics whose genome has been sequenced to find variations), Haplotype (i.e., a set of SNPs that tend to occur and be inherited together and can be linked to a specific disease), AlleleFrequency (i.e., the frequencies of appearance of the alleles of a variation in a specific population), and GenotypeFrequency (i.e., the frequencies of appearance of the genotypes of a variation in a specific population) classes were linked to the SNP class. This is because genomic population studies focused on looking for this type of variation. However, studies that focus on other types of variations are gaining popularity. The CSHG v2 did not represent this a new trend.
2. Variations were modelled redundantly. They were instantiated at least twice: one per frequency and one per type. The rs11571636 variation illustrates this redundancy. This variation was instantiated as a Precise and Indel variation (i.e., by type). It also was instantiated as a Mutant variation (i.e., by frequency). If new genomic studies showing a high frequency of this variation in a specific population are performed, this variations would also be instantiated as SNP. This representation was counter-intuitive and overcomplicated.
3. The definition of exclusive disjunction XOR rules was needed to ensure data correctness. For instance, a variation instantiated as a Polymorphism could only be instantiated as Precise; a variation instantiated as Imprecise, could only be instantiated as Mutant; a variation instantiated as CNV could only be instantiated as Insertion, etc. These rules added an additional layer of complexity to data management.
4. The terminology used was controversial. The mutation and polymorphism terms may lead to incorrect assumptions. A mutation is assumed to have pathogenic effects, while a polymorphism is assumed to have benign effects. Thus, the term variation is preferred over mutation and polymorphism [128].

In the new version of the CSHG (see Fig. 3.3), we use a different approach: the Population, Haplotype, AlleleFrequency, GenotypeFrequency are

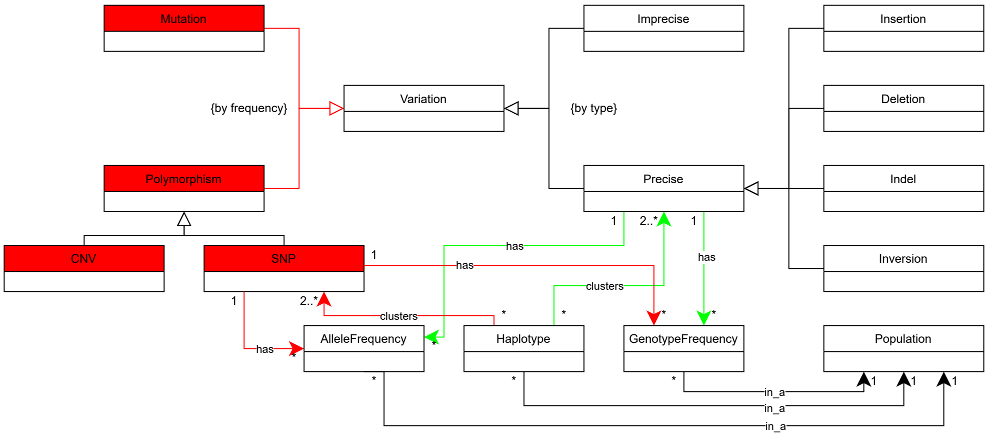

Figure 3.3: Changes for improving the representation of variations. Additions are depicted in green while deletions are depicted in red.

linked to Precise rather than to SNP. This change solves the first issue reported above. Since any Precise variation can have associated information regarding its frequency among populations, the specialisation by frequency is no longer required, and variations are only classified based on their type, which solves the second and third issues. After this update, the former limitations of the CSHG have been overcome, gaining simplicity and expressiveness through an exercise of conceptual reevaluation.

3.1.5 Improving the representation of the effects caused by variations

The CSHG described the effects that variations cause in our body with respect to Phenotype expression alterations (i.e., a high-level perspective). However, these changes are caused by specific changes in the structure of proteins or other genomic elements. The schema represented this by means of variation-phenotype links with a given level of Certainty. The Certainty indicates how strong the evidence supporting a variation-phenotype link is, and it is provided by domain experts as an outcome of their research.

We identified three improvements regarding the effects that variations cause in our bodies:

- The Certainty is provided by domain experts that submit their findings to an authoritative source. However, they were not precisely described in the schema. It is crucial to know who submits evidence and when.
- The CSHG reported variation-phenotype links without indicating the pathogenicity of this link. In other words, does a variation have a pathogenic or a protective effect regarding a phenotype?
- The model did not consider lower-level consequences such as structural changes in proteins.

In the updated model (see Fig. 3.4), we included the notion of Submitter, who report its Submission in a given *Date*. We also added the *clinical_significance* attribute to the Certainty class to represent the concept of pathogenicity. We defined this attribute following the Clinvar recommendations [129], which included the ACMG/AMP recommended terms [128] complemented with additional terms for a more precise characterisation. The most commonly used clinical significance terms include:

1. Benign: a variation that is not responsible for causing a particular phenotype.
2. Pathogenic: a variation that is responsible for causing a particular phenotype.
3. Protective: a variation that decreases the factor of a phenotype.
4. Uncertain: the clinical significance of a variation with respect to a phenotype cannot be assessed.

Finally, we include a low-level representation of the structural changes caused by variations, which was inspired by the standard of the ANN field of the VCF file format [130]. The new Annotation class identifies the *impact* and *effect* of a variation with respect to the ChromosomeElement, Transcript, and Protein concepts. Based on the putative effect of a Variation, there are four impacts:

- HIGH: A disruptive change is triggered. A disruptive change causes a chromosome element truncation or a loss of transcript functionality. For instance, a premature stop codon prevents a protein to from being completed.

- MODERATE: A non-disruptive change is triggered. For instance, an enzyme's sequence is altered, reducing its efficiency.
- LOW: A harmless change is triggered. For instance, a variation changes a CTT codon to CTC, but the protein sequence is not altered because both codons are translated to the same amino acid (Leucine).
- MODIFIER: This is a particular case where the variation is located in a non-coding region. For instance, there is a change in the sequence of an intron region.

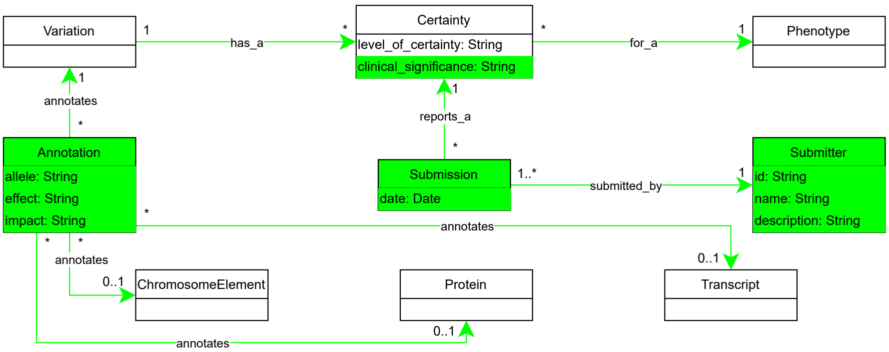

Figure 3.4: Changes for achieving a first approach to model the effects caused by variations. Additions are depicted in green while deletions are depicted in red.

In the new version of the CSHG, the description of the effects caused by variations has been improved. First, it is enriched by including the variation's pathogenicity and the submitters in the CS. Second, a new, low-level approach that focuses on structural changes and its implications at a genome and proteome level is added. As a consequence, the holistic perspective of the CSHG is increased, and the efficiency of domain expert analysis processes is boosted.

3.1.6 Extending the representation of gene products

The characterisation of the protein-coding process in Version 2 of the CSHG was grounded on molecular biology's central dogma [52] (for more information, we refer the reader to page 29), which states that DNA is transcribed into RNA, and RNA is translated into amino acid sequences. The Gene was represented as a DNA element composed of two types of regions: the Exon and the Intron. Only the exons from a Gene were transcribed into a Transcript, which, in turn, was translated into a Protein. Two DNA regions could regulate these processes, namely, the GeneRegulator, and the TranscriptRegulator.

However, three issues that required a precise ontological clarification were identified:

- Genes not only are able to code for proteins, but they can also code for additional products called non-coding RNA (ncRNA) [131]. ncRNAs are associated with a variety of regulatory functions in organs and tissues and are gaining attention in the field of precision medicine [132].
- Those RNA elements that are translated into proteins are called messenger RNA (mRNA), a missing concept in v2 of the CSHG. The reason is that this concept remained implicit inside the Transcript class. However, it was necessary to make it explicit since other types of transcripts exist and must be represented in our schema.
- Regulatory elements were not adequately characterised. First, all of the regulatory elements were represented as DNA elements, even though they also exist at the RNA level. Further, there was no definition of what they were supposed to regulate. For instance, what gene is regulated by what specific enhancer?

Four changes allowed us to provide the required ontological clarification to solve the three issues reported above (see Fig. 3.5). In the first change, we modelled the concepts of mRNA and ncRNA. While mRNA transcripts code for proteins, ncRNA do not (i.e., they remain as RNA sequences in our body). A Gene no longer codes for a Transcript that is translated into a Protein; now, Genes code for either mRNA or ncRNA. Also, we clarified that a Protein is coded from an mRNA rather than a Transcript.

In the second change, we characterise the structural parts of the mRNA. This RNA sequence is composed of three elements: a coding sequence (cds), an untranslated region before the cds called 5' UTR (5utr), and a second untranslated region after the cds called 3' UTR (3utr)

In the third change, we allowed transcripts to be composed not only of exons, but also of introns. Although mRNA is usually composed of exons, there is a mechanism, called intron retention (IR) [133], which is responsible for ensuring the introns are not removed in the transcription process. Also, some ncRNA are obtained from introns. This allows us to represent the following situations:

- The basic protein-coding process in which a transcript composed of exons codes for a protein.
- A typical protein-coding process where, because of IR, an intron is part of the sequence translated into another protein.
- A ncRNA is transcribed from an intron, ready to perform its activities. Most of the ncRNA transcripts are generated from introns.

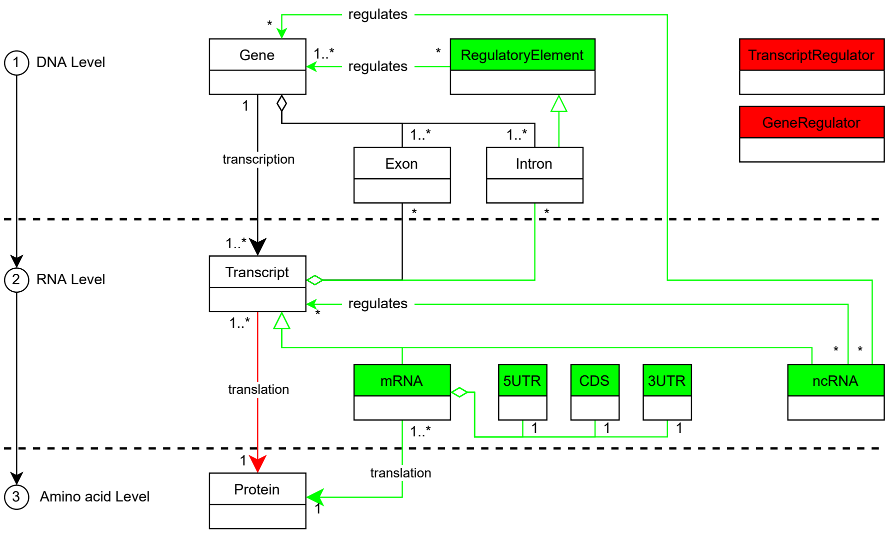

Figure 3.5: Changes as a result of rethinking the gene expression process. Additions are depicted in green while deletions are depicted in red.

In the fourth change, we reevaluate regulatory elements. They are now divided into DNA (the class was renamed to RegulatoryElement, a more appropriate name) and RNA (i.e., ncRNA) regulatory elements. It is important to consider that this division shows at which level they exist, not at what level they act. For instance, ncRNA can regulate the expression of a Gene or a Transcript; but a RegulatoryElement can only regulate the expression of a Gene.

DNA-level regulatory elements are defined as “passive” (i.e., they do not actively regulate gene expression). Instead, they are specific regions where “active” regulatory elements bind. There are four types: promoters, enhancers, silencers, and introns [134]. Promoters are the gene transcription starting point. Enhancers can be bound by activators to boost gene transcription.

Silencers can be bound by repressors to inhibit gene transcription. Introns strongly stimulate mRNA accumulation.

RNA-level regulatory elements are defined as “active” (i.e., they bind to the “passive” elements and actively regulate gene expression). They act at the DNA and RNA levels. While some types of ncRNA bind to silencers to prevent gene expression, others inhibit gene expression by increasing mRNA degradation speed.

3.1.7 Conclusions

Throughout this Section, we have answered **RQ3**: Why/What/How to extend and update the CSHG? It is important to emphasise the value of this first contribution. The extension of the CSHG that has been introduced must not be seen as a simple update. On the contrary, it incorporates complex relevant semantic changes that conform to a much richer version that includes essential conceptual components that were not present in the previous works.

3.2 The Conceptual Schema of the Citrus Genome

As a result of this extension, the CSGH is now potentially ready to be transferred from an academic environment to real-world use cases in precision medicine. At this point, we are ready to tackle the next subgoal, namely, to *explore the conceptualisation of the genome for another species rather than humans* (**G2.2**).

3.2.1 Introduction

The genome is what provides the enormous variability of life that exists on our planet. We have always been interested in providing a holistic perspective of the conceptual modelling efforts oriented to understanding the language of life independently of any particular species. After updating the CSHG, we decided to explore broader domains, which we achieved when an opportunity to collaborate with the *Instituto Valenciano de Investigaciones Agrarias*[4] (IVIA) arose. The IVIA[5] is an internationally-recognised research centre for studying the evolution of the citrus genus and its genome [135].

[4] In English, Valencian Institute of Agricultural Research

[5] https://ivia.gva.es/es/inici

Citrus is a particularly relevant crop that is cultivated worldwide, with a production of more than 100 million tonnes. The Citrus genus comprises more than 1,600 species and includes oranges, lemons, grapefruits, and pummelos, among others. Citrus genome resources are abundant. The first citrus variety was sequenced in 2003 [136], and several genomes have been sequenced since then. More than 67 species have been sequenced multiple times by the time this thesis was being written, with more than 200,000 genes identified[6].

We discovered a significant difference in this working context: domain experts focused much more on technologically-oriented data than purely biological data. For instance, they relied on variant annotations and functional effect prediction software, combining biological and non-biological information. Consequently, citrus data was stored as obtained, mixing biological and technological concepts. In our previous work with the human working context, their maintainers transformed the data that we accessed into a specific model, increasing its abstraction and making it technology-agnostic. However, the citrus data did not undergo this process, and the information was much more tied to the technologies used and their associated limitations. For instance, there was no distinction between qualitative data that indicates the quality of the sequencing process of variants and their biological significance. As a result, domain understanding is affected, being more complex, uncertain, and limited.

One of the main problems that arose was domain heterogeneity, which complicated knowledge extraction processes. We developed the Conceptual Schema of the Citrus Genome (CSCG) to define the final data structure, provide a guide to perform the necessary data transformation operations, and help the IVIA researchers to improve their data management strategies and comparative genomics analyses (see Fig. 3.6). These analyses consist of prioritising variations that are potentially associated with a trait of agricultural interest. In addition, the IVIA uses such analyses to study the evolutionary history of citrus, which is still a matter of controversy [136].

3.2.2 The CSCG

The CSCG is the ontological basis that provides all the necessary information to manage citrus genomic data. Several sessions were needed to implement a first draft of the schema. On the one hand, the IVIA experts provided their vast biological knowledge to understand and interpret the available data correctly. This knowledge allowed us to successfully transform an immense amount of non-structured data into data following a well-defined conceptual schema to

[6] https://www.citrusgenomedb.org/data_overview/1

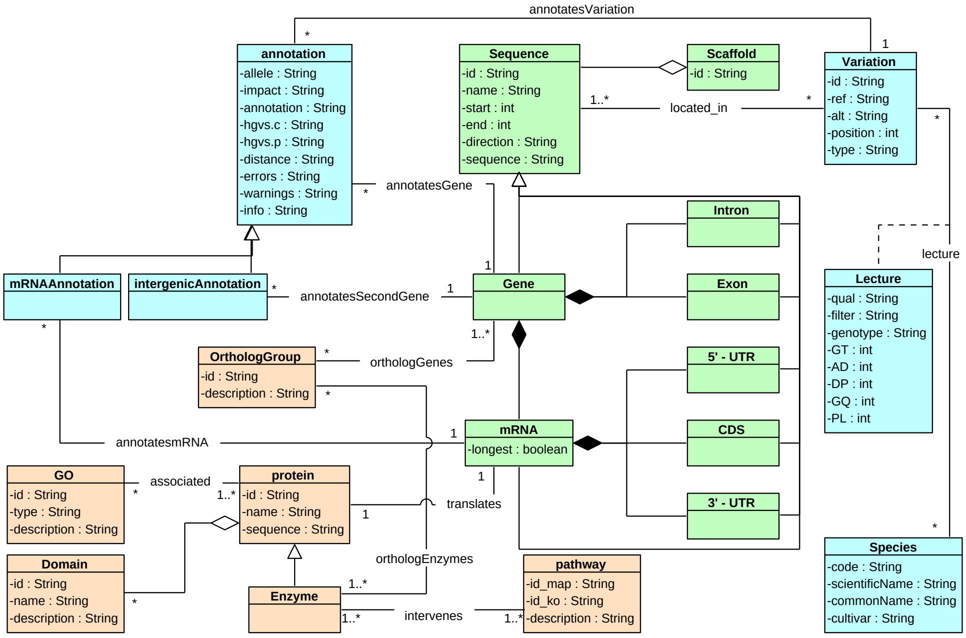

Figure 3.6: The Conceptual Schema of the Citrus Genome. The structural view is depicted in green, the functional view is depicted in orange, and the variations view is depicted in blue.

extract knowledge. On the other hand, the author of this thesis provided its experience in CM to design and implement the conceptual model accurately.

Throughout these sessions, we analysed the data to identify which elements were of higher importance, which allowed us to create and expand the CSCG through an iterative process until a stable version was achieved. Next, we introduce the resulting schema, a precise representation of the genomic domain tailored to the specific needs of the IVIA. The classes that compose the CSCG are grouped into three main views:

- The structural view: This view describes the different regions that can be identified in the DNA sequence of citrus, providing a hierarchical dependency between the regions (i.e., CDSs are located in mRNA, and an mRNA is located in a gene).
- The functional view: This view contains entities with a given function inside the citrus fruits under analysis. These concepts include gene products

(i.e., proteins and their structure), biological pathways, and orthologous groups. This view aims to provide information about how gene products interact with the specific organism effectively.

- The variations view: This last view models the variations that can occur with respect to the reference sequence, the appearance of such changes in specific citrus fruits, and their predicted effect using software annotation tools. [137].

It is worth mentioning that there is one particular aspect that strongly shapes all of their work: because of technological and economic limitations, they use a single sequence of reference that is shared among all of the citrus varieties, namely, the reference sequence of the orange (*Citrus sinensis*). This means that, while each citrus should have its own whole-genome sequence of reference in normal conditions, they only work with the set of variations obtained from comparing their sequence to the reference sequence. This approach saves time and storage because generating a reference sequence is particularly expensive and time-consuming. Also, a file containing the set of identified variations is much smaller and easy to work with than a file containing a whole DNA sequence.

The aspect mentioned above means that only one reference sequence is represented in the model by means of the technologically-oriented Scaffold class, which is a composition of hundreds of small sequences with gaps of known length that have been joined (to simplify, the reader can assume that scaffolds are equivalent to chromosomes). The sequences of reference contained in the scaffolds were obtained from format-specific raw data generated from sequencing machines without undergoing any additional data transformation.

The semantics of the sequences that compose scaffolds have been captured in the schema with the Sequence class. This class is generic enough to allow us to define any arbitrary sequence inside the scaffold and give them semantic meaning. The CSCG represents seven types of sequences, namely, Gene, mRNA, Exon, Intron, 5'-utr, cds, and 3'-utr. Since mRNA and its parts (i.e., the UTR and CDS regions) are transcribed from DNA, then mRNA, 3'-utr, 5'-utr, and cds classes represent the locations in the DNA that are transcribed to obtain them. Again, this representation was a consequence of having to deal with format-specific raw data where RNA-based genomic components were located in a DNA-based system of coordinates.

Genes are the templates used by the cellular machinery to synthesise Proteins, which are composed of Domains. A Domain is a structural part of a

PROTEIN that is common among multiple proteins and is self-stabilising (i.e., a DOMAIN folds independently from the rest of the protein structure). Evolution uses DOMAINS as building blocks to create novel PROTEINS. Every PROTEIN is made of several DOMAINS, but we cannot make such an assertion in the CSCG because current knowledge is incomplete, and there are several PROTEINS whose DOMAINS have not been identified due to technological limitations.

PROTEINS can be associated with Gene Ontology (GO) terms, depicted in the CSCG with the GO class. This class is a technologically-oriented class used to represent the information stored in the GO data source. Most of the information in GO has been generated through computational methods and is very technology-dependent. It is important to note that GO characterises the functionality of gene products rather than genes. Although there are other gene products, PROTEINS are the only gene product that is relevant for this use case. There is a particular type of PROTEIN, known as an ENZYME, whose role is to catalyse (i.e., accelerate) chemical processes. In our schema, these chemical processes are PATHWAYS consisting of several chemical reactions chained together. We have abstracted such complexity because domain experts are only interested in knowing in which specific PATHWAYS ENZYMES work.

There are families of genes that have evolved from a common ancestor; such groups of genes are called orthologous groups (represented in the CSCG through the ORTHOLOGGROUP class). This knowledge is crucial to unravelling the phylogenetic history of citrus, especially if the grouped GENES code for ENZYMES. This is because their changes tend to be more disruptive and, therefore, alter phenotype expression. The available information regarding orthologous groups was stored in raw tabular files that associated the data source-specific identifiers of genes. These genes needed to be mapped and transformed so that they could be integrated with the rest of the data.

Thousands of VARIATIONS in the DNA sequences of citrus specimens are found, and each specimen pertain to a particular SPECIES. Since the same VARIATION can appear in different specimens, the identification of a given variation in a specific specimen is reported through a class called LECTURE, which contained additional specimen-specific information. The data obtained from the sequencing machines was stored in VCF files, combining biological (e.g., the *genotype*) and technological knowledge (e.g., *quality* of the sequencing process) using a technological-specific data format. In addition, VARIATIONS were annotated with functional prediction software. In the schema, ANNOTATIONS indicate structural and functional changes of a variation for a GENE, an intergenic region, or an MRNA depending on its location in the DNA sequence. Since the reference sequence was used for identifying and annotating the DNA variations, the data had a high degree of redundancy. For instance, two VCF files containing the same variation stored the same variation-specific information and SnpEff annotation.

3.2.3 Conclusions

Throughout this section, we have answered **RQ4**: Why/What/How to generate a conceptual schema of the genome for anon-human species? The conceptual schema (i.e., the CSCG) described above constitutes the second contribution of this thesis. Working with such technologically-oriented data required further conceptualisation efforts before carrying out data integration processes. The generated conceptual schema was biased due to the inherent limitations of the data. Even conceptualisation processes were more complicated because domain experts used technology-specific terms rather than their biological equivalents.

The significant differences with the data management of the citrus use case resulted in a conceptual schema that represented some of the concepts very differently when compared to the CSHG representation. However, the CSHG and the CSCG are two instances of the same ontology, namely, genomics. While the CSHG consisted of a precise representation of genomics components, the CSCG included technological aspects because technological and biological concepts were mixed in the data.

3.3 The Conceptual Schema of the Genome

The creation of the CSCG resulted in two conceptual schemas that, in reality, are different instances (i.e., humans and citrus) of the same conceptualization process: the genome. It is clear that differences between both conceptual schemes exist, but they also have strong similarities, such as the division of the DNA sequence into its structural parts or the protein synthesis process. Thus, we propose that it does not matter the species from which the genome is being studied; we should be able to conduct these studies using the same conceptual schema. To achieve this challenge, we need to *generate a conceptual schema that is species-independent* (**G2.3**).

3.3.1 Introduction

The heterogeneity and diversity of genomics use cases we have encountered have been remarkable; the motivating factor for the development of each conceptual schema was quite specific. The CSHG focused on modelling the human genome to improve data management in the field of precision medicine and also sought to improve genetic diagnosing by providing a detailed and holistic representation of the genome. The CSCG focused on modelling how current technologies capture and represent the citrus genome; its goal was to facilitate the identification of DNA variations responsible for the existing variability in the expression of characteristics of agricultural interest.

Despite having developed different conceptual schemes for different species, genome representation is a problem that affects all species of living beings because the genome is what explains life on our planet. Conceptual schemes that focus on a specific species could be seen as a limitation in this context because studying a different species would require creating a new conceptual schema to illustrate the particularities of that species accurately. This is what happens in current practice, thus making it extremely difficult to apply a holistic perspective to the problem of understanding the genome, enabling the understanding of life. This incredible challenge is behind all the work done in this context.

We strongly believe that a conceptual schema that is generic enough to abstract species concepts and conceptually characterises the genome as a whole is a desirable artifact that should be generated and used. To generate such a schema, we began by ontologically compare our two conceptual schemes, namely, the CSHG and the CSCG. From this comparison, we identified their similarities and differences and created the first version of a species-independent conceptual schema (i.e., the CSG).

After generating the CSG, we validated its capabilities with clinical partners that work in a precision medicine context. They identified a set of limitations to be addressed prior to the CSG's use in real-world cases. Overcoming such limitations led to Version 2 of the CSG.

3.3.2 CSG Version 1

The identification of the genome information items that are more relevant for each use case has been an essential task in achieving our goal. The changes required for creating the CSG are analysed and reported below. The comparison has been divided based on the conceptual views of the CSHG because it covers a wider range of relevant genomics concepts. We followed the subsequent order: **i)** the structural view, **ii)** the transcription view, **iii)** the variation view, **iv)** the pathway view, **v)** the bibliography and databank view.

The Structural View

For humans, the CSHG has an abstraction mechanism in which any existing element located in the genome sequence, such as genes, can be modelled. This approach is also used for citrus in the CSCG, meaning that the semantics of the ChromosomeElement and Sequence classes are equivalent. This characterisation is generic enough to achieve our goal and allows for representing any eventual species-specific genomic element. However, the concept of scaffold was not explicitly defined in the CSHG.

As we mentioned above, the concept of scaffold results from the current limitations of current sequencing technologies. Each sequenced species has a set of scaffolds from which chromosome sequences were built. This concept is an example of how core concept definitions can be ambiguous and hard to model even with the help of domain experts. The characterisation of the scaffold was complex. When domain experts used the CSHG, they did not differentiate between a chromosome and its corresponding scaffolds. However, the domain experts that used the CSCG were able to explicitly distinguish these concepts. Indeed, they are not equivalent; the scaffold concept pertains to a technological dimension and will, by definition, represent a fragment of the sequence of a chromosome, which pertains to a purely biological dimension.

We decided to make the Scaffold explicit in the CSG (see Fig. 3.7) because the scaffolding process is inherent to the sequencing process of any species. It is modelled as an entity containing a chromosome sequence. This solution allows us to represent scaffolds containing entire chromosome sequences and scaffolds containing just a part of the entire chromosome sequence. Also, this addition was needed because working with particular scaffolds, rather than working directly with chromosomes, was an essential task in some genomic domains that we had not yet faced[138].

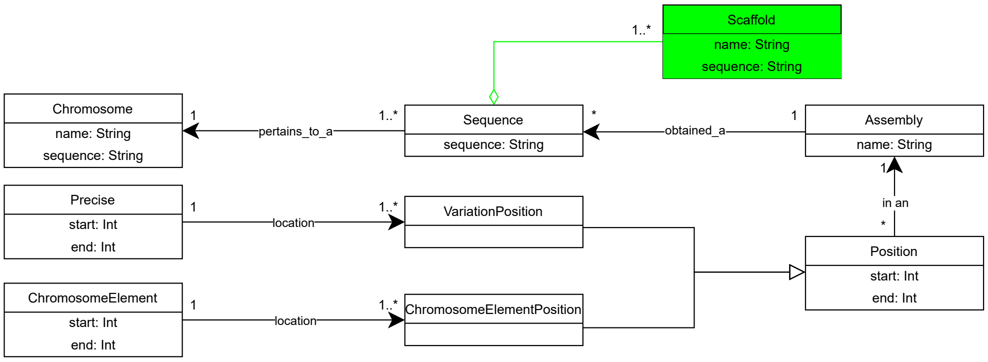

Figure 3.7: Changes for obtaining the structural view of the CSG. Additions are depicted in green while deletions are depicted in red.

The Transcription View

The CSHG and the CSCG have a high degree of similarity regarding the representation of transcription processes. The CSHG differentiates between those ChromosomeRegions that can be transcribed (i.e., TranscriptableElements) and those that regulate such transcription (i.e., RegulatoryElements). The most relevant TranscriptableElements element is the Gene, which we defined as "a union of genomic sequences encoding a coherent set of potentially overlapping functional products" [139]. This definition is represented in the CSHG using an aggregation relationship between the Gene and the ChromosomeElement. This approach also allows us to instantiate more exotic occurrences such as genes with regulatory elements of other genes inside their sequence, nested genes (i.e., a gene that inside a larger gene) [140], or trans-splicing (i.e., a transcript originated from different gene sequences) [141].

The main difference between transcription in the CSHG and in the CSCG is that the latter is not able represent regulatory elements because, unlike with humans, the knowledge needed to identify them in the citrus genome is not yet achieved. However, the CSCG has two concepts that are missing in the CSHG, namely, the protein Domains and the groups of orthologous genes (represented in the CSCG as OrthologGroups).

Protein domains are defined as "the basic, independent unit of protein folding, evolution, and function" [142]. The CSHG represents proteins as a homogeneous unity that performs a given functionality. However, each protein domain performs a specific task contributing to protein functionality. An orthologous group is a set of genes that have evolved from a single gene in a common ancestral and are created by speciation events.

We included these two missing concepts in the transcription view of the global CSG (see Fig. 3.8). With these additions, proteins are composed of DOMAINS, and the ORTHOLOGGROUP clusters sets of genes and, optionally, another set of associated proteins. As a result, we can decompose proteins into their molecular domains, which is essential for providing accurate protein functional classification [143]. We can also apply phylogenetic studies and comparative analysis to study thousands of evolutionary studies and improve gene identification [144].

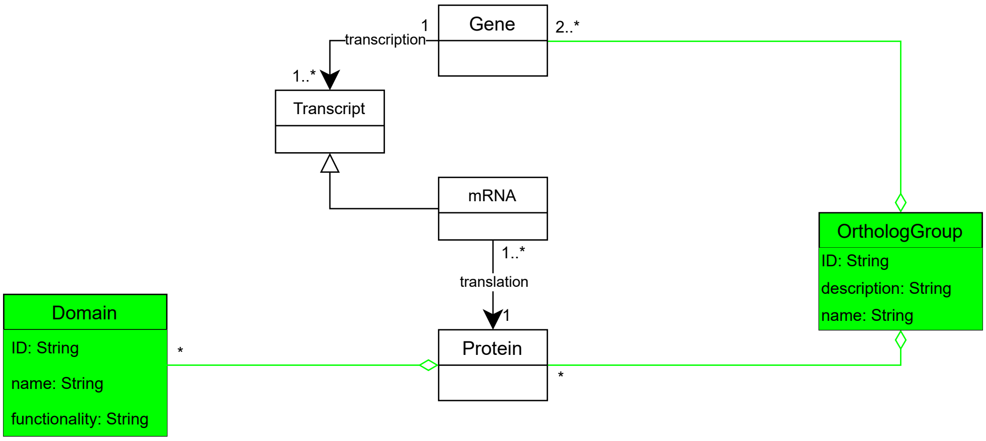

Figure 3.8: Changes for obtaining the transcription view of the CSG. Additions are depicted in green while deletions are depicted in red.

The Variation View

The CSHG was conceived as an abstract representation of the genome itself. This means that, unlike the CSCG, it does not represent individuals. Thus, we identified two limitations of the variation view of the CSHG compared to the CSCG. The first limitation was the absence of the SPECIES concept, which is crucial for a species-independent conceptual schema. The CSCG considered this concept because domain experts worked with several citrus species.

The second limitation is how variations are represented, which is related to not representing individuals explicitly. The CSHG modelled variations at the general level (i.e., the variation itself) and at the population level (i.e., the frequency of appearance of a variation in a given group of individuals); but it did not model the individual level (i.e., individual-specific particularities of a variation). For citrus, the CSCG considered variations at the general and individual levels, but it did not consider the population level.

To overcome these limitations, we first included the concept of Species in the CSG (see Fig. 3.9); we defined it as an entity associated with its corresponding set of chromosomes. Then, we modelled the individual level by creating the notion of Individual and connected it to the Population class to show that a Population is composed of a set of Individuals. Finally, we added the Lecture entity as an association class between Individual and Variation in order to allow the individual level in the CSCG.

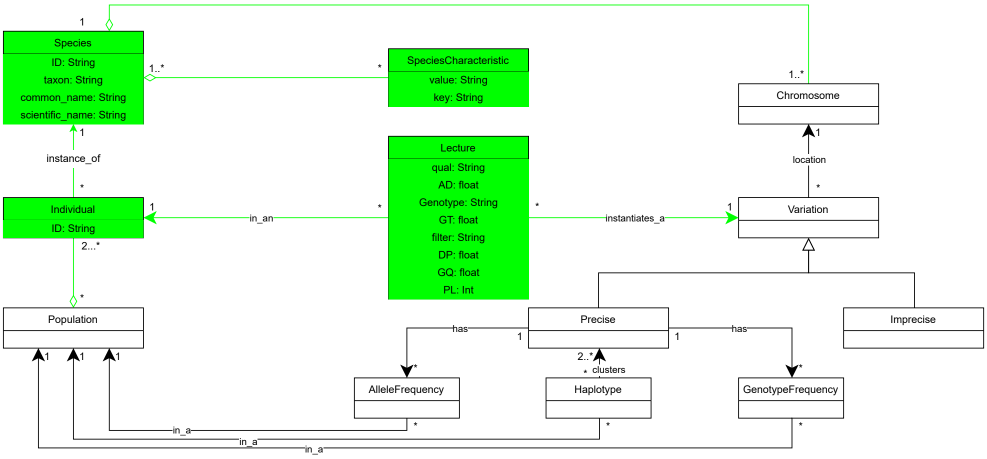

Figure 3.9: Changes for obtaining the variation view of the CSG. Additions are depicted in green while deletions are depicted in red.

The update described above resulted in the improvement of management and knowledge extraction for both human and citrus data. For humans, we can study the appearance of variations in specific individuals. For citrus, we can focus on the study of haplotypes, which is a topic of interest to estimate heterozygosity rates of citrus species [145].

The Pathway View

The CSHG represented pathways and their inner processes with a highly generic and flexible approach, including the events that compose each pathway, how they are related, and the participating biological entities. This approach allows us to model pathways with the desired level of granularity and generate hierarchical structures with them. The upper-level concept of the event specialises into the process (i.e., an atomic event that cannot be further decomposed) and the pathway (i.e., complex events that can be further decomposed into other events, either pathways or processes). Each process (i.e., an atomic event) is associated with the set of biological entities that act as input, output, or regulator.

The level of knowledge associated with biological pathways in citrus is much lower. This knowledge gap can be seen in the CSCG, where pathways are represented as indivisible events, ignoring their internal processes. Also, only enzymes are associated with pathways without considering other relevant biological entities participating in pathways. We conclude that no changes to this view of the CSHG are needed.

The Bibliography and Databank View

The CSHG allowed us to model the bibliography of the represented genomics components and identify them in external data sources. The CSCG did not include information regarding the bibliography of the data or its origin. Therefore, no changes to this view of the CSHG are needed.

Conclusions

Our experience showed that working with different species leads to focusing on different dimensions. The result of our work led to the generation of the first version of a species-independent conceptual schema (i.e., the CSG). This result responds to **RQ5** and is the third contribution of this thesis. The CSG was developed by comparing the artifacts generated in two individual exercises in which conceptual modeling techniques were applied to represent the genome of two different species (i.e., human and citrus). This comparison highlights the importance of explicitly differentiating technological and biological aspects when modelling the genome. It also shows the main conceptual differences between characterising the genome and its manipulation as a generic model versus its particular instantiation in individuals.

The experience accumulated from analysing the human and citrus genomes has allowed us to design a holistic conceptual schema that captures the essential aspects of the genome structure, identifying all of the relevant concepts that represent the knowledge associated with the genome regardless of the species.

An opportunity to validate the capabilities of the CSG together with real-world users arose via a collaboration with clinical partners, discussing the CSG, who focus on precision medicine and genetic diagnosis. As a result of the preliminary interactions with the new partners, we identified specific parts of the schema that required further attention and extension before its use by real-world users. The next section presents a comprehensive discussion of the identified limitations and how they were resolved.

3.3.3 CSG Version 2

The emergence of precision medicine has transformed the understanding of medicine, moving from a reactive approach focused on curing diseases, towards being more proactive and aimed at disease prevention. Apart from prevention, precision medicine also aims to improve diagnosis and treatment by providing individualised treatments. Although the concept of precision medicine is not clearly defined [146], some attempts have been made at its delimitation. König et al. define precision medicine as “a standardized process that incorporates and exploits clinical data, lifestyle, or genomics information, among others” [146], while Agusti et al. define it as “the stratification of patients using novel approaches” (i.e., genotype-based approaches rather than typical symptoms) [147].

Providing high-quality precision medicine depends not only on genomics data but also on other omics data such as proteomics or metabolomics. Based on the definitions of precision medicine provided above and the discussions with domain experts, we have identified four issues that need resolving to improve the CSG:

1. **Proteomics**[7]: The first issue is that the representation of proteins lacked depth in Version 1 of the CSG. Proteins were represented as biological entities composed of smaller building blocks called domains. However, in proteomics, many more pieces of information that were not explicitly represented (e.g., the protein’s size, structure, or physicochemical properties) are not taken into account. In order to exploit the full potential

[7] A field dedicated to the large-scale study of proteins.

of proteomics and increase the accuracy of precision medicine, the CSG must extend and improve how proteins are represented.

2. **Clinical Actionability**: Precision medicine is still based on interpreting or predicting the consequences of DNA variations in the human body, which are summarised using the concept of “clinical significance”. In this context, correctly assessing the clinical significance is crucial. However, submitters interpret DNA variations several times and for different phenotypes, and the information they provide differs frequently. Therefore, the CSG should better capture possible discrepancies regarding the clinical significance of variations.

3. **Biological Entities and Metabolomics**[8]: After a first approach to studying metabolomics, it was noted that the representation of biological entities and how they are interconnected needed to be improved. In addition, the classification of biological entities was not ontologically clear.

4. **The role of biological entities**: The last improvement was related to the role biological entities play with respect to a phenotype. What function of which protein is altered when a genetic-based disease is manifested? How are protein or enhancers' functions altered when DNA variations occur? Answering these questions was impossible with Version 1 of the CSG, which needed to represent the roles of biological entities and how variations can alter them.

We elaborate on each of these four problems through the rest of this section. Solving them has led us to the current, last version of the CSG: Version 2.

Proteomics

Proteins are macromolecules that play a fundamental role within every cell and metabolic reaction. For instance, protein kinases are associated with learning and memory processes [148], while high concentrations of C-reactive protein (CRP) increases the risk of developing heart diseases [149], etc. The first version of the CSG lacked appropriate protein representation and, thus, required the addition new concepts and relationships to cover three missing topics:

- Describing protein properties [150], [151]: proteins are not isolated entities, they have biophysical and chemical properties that interact with

[8] A field dedicated to studying metabolites and their interactions in cell chemical processes.

the surrounding environment. This alters their functionality and efficiency. For instance, the optimum pH for the Phosphatidylserine lipase (ABHD16A) enzyme is between 7.2 and 8.0 [152], the Pyruvate kinase (PKM) enzyme has a Michaelis constant (KM)[9] value of 2.7 mM for phosphoenolpyruvate at 32 degrees Celsius [153], etc.

- Characterising protein isoforms [154]: protein isoforms are highly similar proteins that are obtained from the same gene or family of genes. They can be generated by either alternative promoter usage [155], alternative splicing [156], alternative initiation [157], or ribosomal frameshifting [158]. It is crucial to consider not just proteins but also their isoforms since they can have different functionalities. For instance, the Epidermal Growth Factor Receptor (EGFR) protein has four isoforms. While isoforms 1, 3, and 4 are growth factors, isoform 2 acts as an antagonist. Also, isoform 2 is much smaller [159] and is expressed in ovarian cancer [160]. Another example is the Vascular Endothelial Growth Factor A (VEGFA) protein, which has seventeen isoforms.
- Representing protein structure [161]: proteins fold in space, creating three-dimensional structures. These structures are stabilised by means of polar hydrophilic hydrogen, ionic bond interactions, and internal hydrophobic interactions between non-polar amino acid side chains [162]. Protein functionality depends on its three-dimensional structure; thus, studying protein structure can improve identifying the potential functionality of novel proteins and infer how to change such functionality by altering the protein sequence. Since the availability of new protein sequence data continues to outpace the availability for generating experimental protein structure data by far, there is a great need for accurate protein modelling tools [163].

The changes that allowed us to deal with the three topics described above are depicted in Fig 3.10. We started by studying the currently available knowledge in the domain by performing an in-depth study of the Universal Protein Resource (UniProt) database [164], which stores valuable information regarding protein-related knowledge. After gathering all of the relevant information, we discussed the three topics mentioned above.

For describing protein properties, we included the most relevant Protein and Enzyme properties. For Proteins, we considered the following properties:

[9] The KM constant indicates the rate attained when the site of an enzyme saturates with a substrate.

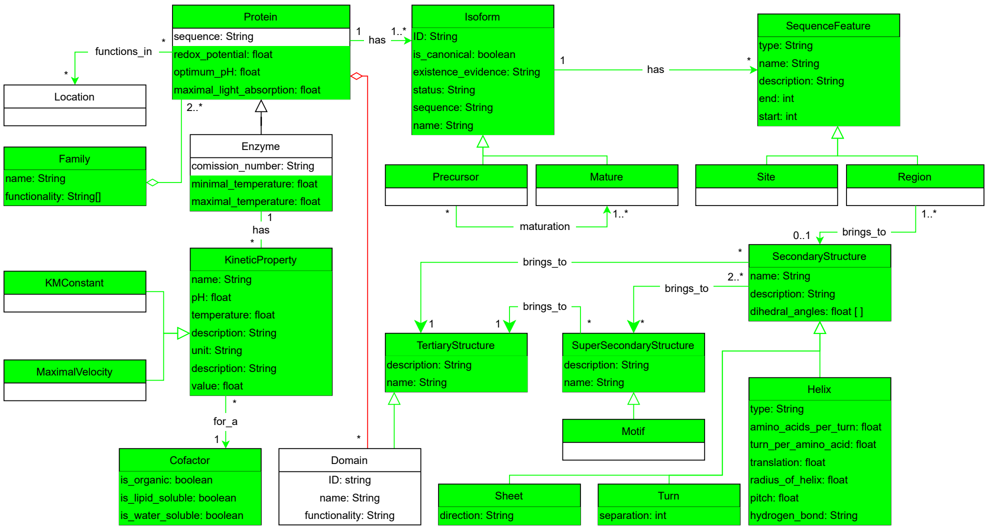

Figure 3.10: Changes for obtaining an extended and improved representation of proteins. Additions are depicted in green while deletions are depicted in red.

- *Redox potential*: the tendency of a PROTEIN (in mV) to gain or lose electrons.
- *Optimum pH*: the pH at which PROTEIN activity is more efficient.
- *Maximal light absorption*: the wavelength (in nm) at which photo-reactive PROTEINS show their maximal light absorption.

For ENZYMES, we considered the following properties:

- *Minimal temperature*: the minimum temperature an ENZYME requires to perform its activity.
- *Maximal temperature*: the maximum temperature an ENZYME requires to perform its activity.
- *KM constant*: the substrate concentration at which half of the ENZYME's active sites are occupied by a COFACTOR (i.e., it measures the affinity of an ENZYME for a substrate).
- *Maximal velocity*: the substrate concentration at which the ENZYME's active sites are occupied by a COFACTOR.

The *KM constant* and the *Maximal velocity* measure the affinity of an Enzyme for a Cofactor. They can be organic or non-organic, water-soluble, and lipid-soluble.

Proteins can be grouped into Families based on their *functionality*. We also represented the Locations where they act.

For characterising protein isoforms, we modelled that each Protein is associated with a set of Isoforms, from which one is *canonical* (i.e., the most prevalent Isoform, which contains the consensus *sequence* of a given Protein). New protein Isoforms can be identified over time with experiments or computational inference. Based on how an isoform has been identified, we give them a *level of evidence*. Besides, isoforms can be Precursor or Mature. Precursor isoforms undergo post-translational processes to bring the final, functional protein (i.e., Mature).

To represent protein structure, we described the arrangement of proteins in space through the primary, secondary, and tertiary structures. The primary structure defines the *sequence* of an Isoform. For the *canonical* Isoform, the *sequence* has a set of SequenceFeatures, which are specific regions of the *sequence* considered of interest. Sites describe single amino acid sequences such as cleavage, inhibitory, or breakpoint sites. Regions describe sequences of more than one amino acid with a functional or biological interest such as a region that mediates transcriptional activity.

The Regions identified in the primary structure fold and stabilise in three-dimensional local segments called SecondaryStructures [165]. These structures fold with specific *dihedral angles* [10]. Here, the ϕ and the ψ angles are used. They have a limited number of possible values due to the existing chemical effects inside the protein structure (the possible values are identified in Ramachandran plots). The most common SecondaryStructures are the Sheet (i.e., the amino acid sequence has an almost linear structure), the Turn (i.e., the amino acid sequence changes its direction), and the Helix (i.e., the amino acid sequence is arranged in a spiral).

Helices are the most complex SecondaryStructures. They are characterised by its *type*, *amino acids per turn*, *translation*, *radius*, *pitch*, and *hydrogen bond*. The *type* indicates whether it is right-handed or left-handed. The value for *amino acids per turn* indicates the number of amino acids needed per turn of helix, and the *turn per amino acid* (i.e., the turn in degrees caused per

[10] A dihedral angle is the internal angle of an amino acid sequence at which two adjacent planes meet.

Name	Dihedral angles	Type	Aa per turn	Turn per Aa	Translation	Radius of helix	Pitch	Hydrogen bond
Alpha	[-60, -45]	R/L	3.6	100º	.15	.23	.54	i + 4 -> i
310	[-49, 26]	R/L	3	120º	.2	.19	.06	i + 3 -> i
pi	[55, 70]	R/L	4.4	87º	.11	.28	.48	i + 5 -> i

Table 3.1: Example of helix structures. Aa refers to amino acid. For simplicity, the description field is not shown. R/L means that the helix can be either right-handed or left-handed.

amino acid) is derived from this value. The *translation* indicates the translation distance (in nm) per amino acid. The *radius* indicates the radius (in nm), of the helix. The *pitch* indicates the vertical distance (in nm) between two turns. Finally, the *hydrogen* bond indicates the type of hydrogen bond of the helix. Table 3.1 illustrates some examples of HELICES.

Multiple SECONDARYSTRUCTURES can be grouped to form a SUPERSECONDARY STRUCTURE [165]. For instance, a β-barrel is composed of a tandem of β-sheets; a helix hairpin is composed of two antiparallel α-helices; a β-hairpin is composed of two β strands connected by a loop. There is a particular type of SUPERSECONDARYSTRUCTURE called MOTIF, which appear in the sequences of several evolutionarily unrelated proteins.

Finally, we modelled the TERTIARYSTRUCTURE elements of the PROTEIN, which are generated when the SECONDARY and SUPERSECONDARY structures fold together and constitute the whole three-dimensional arrangement in space [165]. There is a particular type of TERTIARYSTRUCTURE called DOMAIN, which already existed in the first version of the CSG. DOMAINS are the building blocks of the proteins. They are self-stabilising regions that fold independently and have specific functionality. A large number of DOMAINS have been identified, each with a specific function [166].

Even though MOTIFS and DOMAINS are both composed of SECONDARYSTRUCTURES, they are completely different nature. MOTIFS are assembled by the connection of HELICES and SHEETS through TURNS; they have a structural function and are not independently stable. DOMAINS, on the other hand, can be assembled by SECONDARYSTRUCTURES and SUPERSECONDARYSTRUCTURES structures and use disulfide bridges, ionic bonds, and hydrogen bonds. They have a unique function are independently stable.

The changes herein presented allowed us to represent the three missing topics. First, proteins' most relevant biophysical and chemical properties have been identified. Second, protein isoforms and the existence of precursor proteins that maturate into their final form are now considered with a high level of detail. Third, the three-dimensional arrangement of proteins in space have been characterised (i.e., the primary, secondary, and tertiary structures). These changes also allow us to identify variations that alter proteins' structural and functional units in the three defined dimensions.

Clinical Actionability

The clinical significance associated with DNA variations (i.e., their consequences in our body with respect to a phenotype) is a crucial aspect of precision medicine. Interpreting the clinical significance of a variation is a challenging process that requires gathering and assessing the available evidence. Standards and guidelines such as the ACMG/AMP [17] or Sherloc [167] guide this process and improve knowledge generation. The first version of the CSG associated VARIATIONS with PHENOTYPES through the CERTAINTY class, which described the *clinical significance* and the *certainty* of such association. The *certainty* represented the relevance of the evidence used to establish the association. It is crucial in genetic diagnosis since it provides clinicians with a means to determine whether to include or discard variations when performing genetic diagnosis.

Several publicly available databases, such as ClinVar, Ensembl, ClinGen, or CIViC, provide clinicians and geneticists with thousands of variations with their corresponding clinical significance(s). However, this clinical significance is usually reported as a unique value without considering that it can change from one phenotype to another. Even worse, different data sources can suggest different results. This leads to misinterpretations, worsening precision medicine diagnosis and introducing an unbearable lightness in such an essential interpretation [168]. More specifically, we identified the following two problems:

- The clinical significance of variations is usually reported as a whole rather than at the phenotype level. This means that if a variation is pathogenic for a given phenotype but benign for another, it will be reported as a variation with conflicting interpretations, which is incorrect because there is no such conflict. Thus, the clinical significance must be reported at a phenotype level.
- Because of the situation reported above, conflicts between different interpretations for a variation are not managed appropriately, leading to an imprecise and deficient result.

Clinical experts should review each interpretation to assess the correct clinical significance of variations. This process (if carried out) is tedious, manual, error-prone, and diminishes the added value of this information for precision medicine. To solve the above-mentioned problems, we started by precisely characterising the existing clinical significance types. We identified thirteen possible values for the clinical significance of a variation:

1. **Pathogenic**: increases the susceptibility of predisposition to a certain Mendelian disorder.
2. **Benign**: reduces the susceptibility of predisposition to a certain Mendelian disorder.
3. **Likely Benign**: strong evidence in favor of reducing the susceptibility of predisposition to a certain Mendelian disorder.
4. **Likely Pathogenic**: strong evidence in favor of increasing the susceptibility of predisposition to a certain Mendelian disorder.
5. **Affects**: causes a non-disease phenotype, such as lactose intolerance.
6. **Drug Response**: alters a specific drug response in some way.
7. **Confers Sensitivity**: confers sensitivity to a specific drug.
8. **Association**: identified the association to a disorder in a GWAS study.
9. **Uncertain Significance**: limited evidence regarding pathogenicity.
10. **Protective**: decrease the risk of suffering from a disorder.
11. **Conflicting data from submitters**: groups within a consortium have conflicting interpretations of a variation.
12. **Not Provided**: no clinical significance reported.
13. **Other**: any other possible value.

All these values can be grouped according to their likelihood to cause a potentially damaging phenotype. Thus, we created the ClinicalActionability concept, which is obtained from aggregating all of the *clinical significances* in a Variation-Phenotype association. The ClinicalActionability concept provides a more precise assessment of the actual Variation *clinical significance* on a per-phenotype basis.

Fig. 3.11 shows the changes made in Version 2 of the CSG to obtain a better representation of the concept of clinical significance. We included the ClinicalActionability class, which is associated to a Variation, a Phenotype, and the Certainties from which the ClinicalActionability *value* has been calculated. There is only one ClinicalActionability for a Variation-Phenotype association.

The *value* attribute of the clinicalActionability class has five possible values:

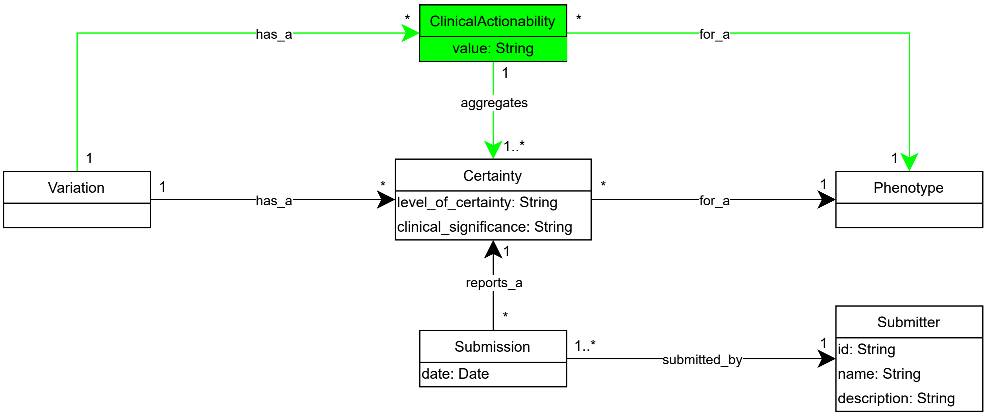

Figure 3.11: Changes for obtaining a better representation of the clinical significance. Additions are depicted in green while deletions are depicted in red.

1. **Disorder causing or risk factor**: The VARIATION causes a phenotype or increases its likelihood of appearing. This value considers the following *clinical significances*: pathogenic, likely pathogenic, affects, risk factor, or association.

2. **Uncertain role**: The role of the VARIATION in the development of a phenotype is not clear. This value considers the "uncertain" *clinical significances* or when there are conflicts between *clinical significances*.

3. **Not disorder causing or protective effect**: The VARIATION does not cause a the phenotype or it provides a protective effect against it. This value considers the following *clinical significances*: benign, likely benign, association not found, or protective.

4. **Affects drugs or treatment response**: The VARIATION affects the sensitivity or response of a drug or treatment. This value considers the following *clinical significances*: drug response or confers sensitivity.

5. **Not provided**: The *clinical significance* of variation is unknown. This value considers the following *clinical significances*: unknown and not provided.

Let us illustrate the benefits of the clinical actionability. The c.2843G>A variation [169] is reported in ClinVar as a variation with conflicting interpretations.

However, when it is analysed at a phenotype level, we see that there is no conflict regarding its reported clinical significances (see Table 3.2).

Interpretation	**Phenotype**
Pathogenic	Leber congenital amaurosis 8 Retinal dystrophy Retinitis pigmentosa 12 Pigmented paravenous chorioretinal atrophy CRB1-Related Disorders Abnormality of the eye Retinitis pigmentosa
Benign	Pigmented paravenous chorioretinal atrophy

Table 3.2: The list of clinical significances associated to the c.2843G>A variation in ClinVar.

Since variations with conflicting clinical significance are usually discarded [170], the inclusion of the clinical actionability allows for considering some of these discarded variations. We analysed all of the variations reported in ClinVar and found that 41,433 have conflicting clinical significances. Cardiopathies, cancer, and muscular dystrophies are the most affected diseases associated with these variations. The clinical actionability will help clinical experts to filter some of these variations better and avoid missing important information.

Biological Entities and Metabolomics

The conceptualisation of biological entities in the Version 1 of the CSG required an improvement. The previous approach considered four types of entities: SIMPLE, COMPLEX, POLYMER, and ENTITYSET.

SIMPLE entities were the elementary entities that took part in processes. There were six types of SIMPLE entities[11]:

- DNA: represented all of the possible molecules of DNA.
- RNA: represented all of the possible molecules of RNA.
- PROTEIN: represented all of the possible molecules of amino acids.
- BASIC: represented those SIMPLE entities that are not DNA, RNA, or PROTEINS.

[11] For simplicity, Fig. 3.12 does not depict the relationships between the child classes of polymer and monomer (e.g., the relationship between DNA and Nucleotide.

- Nucleotide: represented the atomic elements that compose dna and rna.
- AminoAcide: represented the atomic elements that compose Proteins.

Apart from the Simple entity, there were three other types of entities. First, the Complex, which was an aggregation of at least two Entities (called Components) that contribute differently to the whole. Second, the EntitySet, which clustered Entities that play an equivalent role in a given Process. Third, the Polymer, which was a concatenation of several instances of the same type of Entity.

There were five issues regarding how Entities were characterised:

- EntitySets clustered Entities that can be interchanged in an event. However, they were not associated with that specific event.
- The representation of some of the concepts was confusing because they were represented twice in the model. For instance, the dna is not only a Simple entity, but it is also a Polymer.
- dna, rna, Proteins, Nucleotides, and Aminoacids should not be represented at the same level of hierarchy because the last two are the compositional parts of the first three.
- Only one type of AminoAcid chain was represented: the Protein. However, other types, such as oligopeptides, exist and should be considered.
- Saccharides (i.e., carbohydrates and sugars) were not described in the model. Because of their biological importance and complexity, they should be explicitly represented and characterised.

We solved the first issue by creating a new association between the EntitySet and the Event. Besides, the EntitySet is no longer a type of Entity. Conceptually speaking, an EntitySet groups a set of Entities that can be used indistinctly in a Event because they play an equivalent role, but it is not a sub-type of entity.

Regarding the second, third and fourth issues, we completely reevaluated the concept of Entity. As Fig 3.12 shows, the Entity is now specialised into Complex, Simple, and EntitySet. The concept of Simple is more generic and contains three sub-types, namely, Monomer, Polymer, and basic.

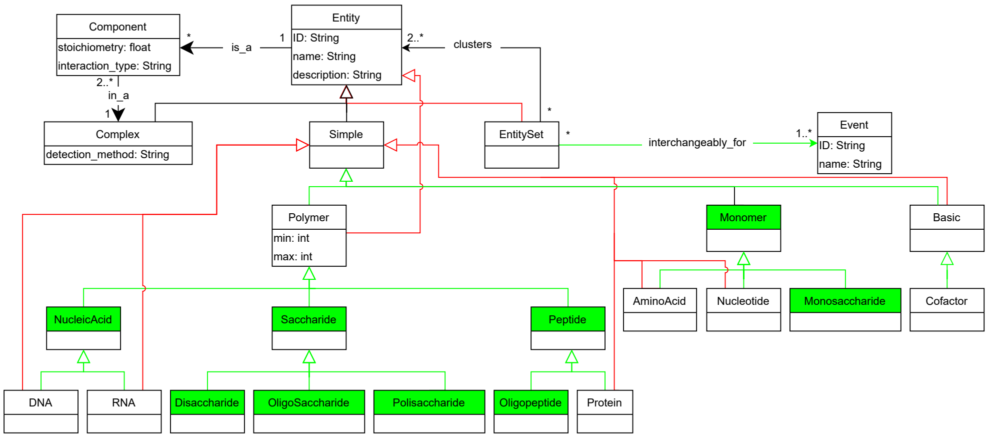

Figure 3.12: Changes for obtaining an improved representation of biological entities. Additions are depicted in green while deletions are depicted in red.

The Monomer represents the atomic molecules that can be chained to create larger entities with homogeneous internal structure (i.e., Polymers). Monomers are characterised by their physical structure (i.e., cyclic or linear), their polarity (i.e., polar or non-polar), and their skeletal formula. The most representative Monomers in life forms are the Nucleotide, the Monosaccharide, and the AminoAcid [171].

The Polymer represents the larger entities that are composed of a set of Monomer entities of the same type. There are three types of Polymers, each with its corresponding Monomer. The first type is the NucleicAcid, which is composed of Nucleotides. A nucleic acid can be DNA if it is double-stranded and contains Thymine, or RNA if it is single-stranded and contains Uracil. The second type is the Saccharide, which is composed of Monosaccharides. A Saccharide is a Disaccharide if it is composed of two Monosaccharides. A Saccharide is an Oligosaccharide if it is composed of between two and twenty Monosaccharides; otherwise, it is a polysaccharide [172]. The third type Peptide, which is composed of AminoAcids. A Peptide is an Oligopeptide if it is composed of less than fifty AminoAcids; otherwise, it is a Protein [173]. All of these restrictions regarding the size such biological entities are captured though the *min* and *max* attributes of the Polymer class.

Lastly, the Basic represents chemical compounds that can be used in biological processes, such as water or Adenosine Triphosphate (ATP). Some Basics can be required by ENZYMES to perform their catalytic activity (i.e., COFACTORS). For instance, Vitamin C is an organic, water-soluble COFACTOR; Vitamin A is an organic, lipid-soluble COFACTOR; and Zinc is an inorganic COFACTOR.

These changes allowed us to correct the five issues mentioned above: ENTITYSETS are associated with the EVENT of interest; we improved the representation of those concepts whose representation was not precise; we made differentiations between POLYMERS and MONOMERS; we expanded the types of AMINOACID chains (i.e., PEPTIDES) represented in the model; finally, we included the SACCHARIDES in the schema.

The Roles of Biological Entities

Every biological entity has a set of specific roles in our body. However, they can be altered when DNA variations are present in the genome, manifesting genetic diseases. This was not adequately represented in Version 1 of the CSG. The main limitations of the first version of the CSG can be summarised into the following four points:

- The high-level role of biological ENTITIES with respect to a PHENOTYPE was not explicitly represented.
- The alteration of such roles caused by VARIATIONS was missing.
- An ENTITY TAKESPART in a PROCESS in a specific LOCATION, not the entire body; this was not considered. Also, it is crucial to distinguish between the location of an ENTITY (e.g., between extracellular glucose and cytosolic glucose).
- We could only represent changes (i.e., PROCESSES) that modify ENTITIES; but we could not represent other changes or effects such as the movement of an ENTITY inside a cell or the change of intracellular pH.

For representing alterations in ENTITYROLES because of DNA VARIATIONS, we specialised this class into the DEFAULTROLE and the ALTEREDROLE. ALTEREDROLES can be caused by a set of VARIATIONS. Here, we decided not to consider a minimum cardinality of 1 because, sometimes, there is no knowledge available regarding which VARIATIONS alter a specific ROLE.

Some examples that demonstrate the importance of such an addition in the CSG are the following:

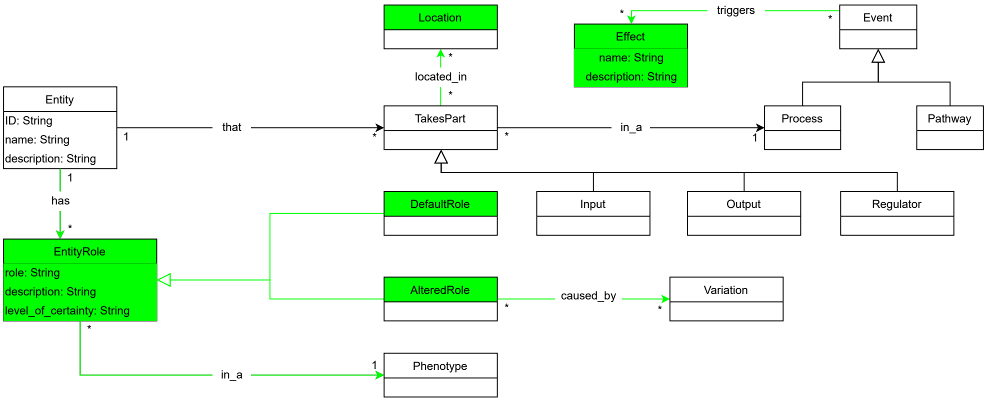

Figure 3.13: Changes for obtaining a better representation of the roles of biological entities. Additions are depicted in green while deletions are depicted in red.

- The CD22 gene codes for the B-cell receptor CD22 protein. The default role of this protein is to disrupt microglia cell function, but an altered role causes detritus accumulation in the brain, which produces Alzheimer's disease [174], [175].
- The Optineurin protein is coded by the OPTN gene, and it plays an important role in the maintenance of the Golgi complex, in membrane trafficking, and in exocytosis [176]. The expression of the Optineurin protein is regulated by intraocular pressure [177]. However, some DNA variations causes the Optineurin protein to selectively promote cell death of the retinal ganglion [178] or it increases the risk of suffering Normal Pressure Glaucoma (NPG) [178].

Further, more examples can be found in several degenerative and neurodegenerative disorders, such as Parkinson's disease [179]. Annotating the Roles of Entities is crucial in establishing genotype-phenotype associations, which is essential for understanding how life works.

For modelling location-associated information, we created the Location class. Initially, we associated this class to the Entity. However, we identified some situations that could not be adequately represented with this approach:

- Entities can have **different** Roles in the **same** Location.
- Entities can have the **same** Role in **different** Locations.

- ENTITIES can have **different** ROLES in the **different** LOCATIONS.
- The ROLE of an ENTITY can **change** depending on the specific PROCESS it is involved in.

Thus, we linked the LOCATION with the TAKESPART class. With this approach, ENTITIES TAKEPART in PROCESSES within specific LOCATIONS, which is a significantly more precise conceptualisation of how life works.

Finally, we observed that, in Version 1 of the CSG, the result of EVENTS only considered OUTPUT ENTITIES. This is a useful and correct approach, but it is not complete because EVENTS can cause a more abstract change in our body. For instance, the "transportation of bicarbonate through the ion channels" EVENT regulates the body's pH. To consider such changes , we created the notion of EFFECT, a more generic consequence, characterised by a *name* and a *description*, that is triggered after a given EVENT occurs.

Conclusions

All the improvements discussed above focused on covering different aspects of precision medicine, resulting in Version 2 of the CSG, a conceptual framework that can be applied in different genomic contexts. This result is associated with **RQ5** and is the fourth contribution of this thesis. This contribution is noteworthy for two reasons. First, it is a sound conceptual representation of the genome that is generic enough to be used regardless of the species. Second, domain experts have reported that our schema is ready to be used in real-world use cases.

3.3.4 Conclusions

Throughout this section, we answered **RQ5**: Why/What/How to generate a conceptual schema of the genome that is species independent? First, we ontologically compared the CSCG and the CSHG to generate the CSG. Then, it was validated by domain experts, who raised concerns that required further improvement of our schema, leading to version 2 of the CSG. For the first time, we have a conceptual modelling artifact that is ready to be used in real-world scenarios.

However, these domain experts also pointed out that the size of the CSG may be too broad for certain specific use cases. The CSG is appropriate for domain understanding and communication, but there may be instances where sometimes the concepts needed for real-world use cases will only cover a small part of the schema. As examples, a detailed characterisation of protein structure is not needed to integrate the clinical significance of variations associated with a given disease, or the characterisation of biological pathways is not necessary when geneticists focus on population studies.

3.4 The ISGE Method

Considering the feedback from domain experts and in order to promote the use of the CSG, it must adapt to the particularities of each use case and provide only the relevant pieces of information needed. Thus, we must *facilitate the adoption of the CSG to work with genomics use cases* (**G2.4**).

3.4.1 Introduction

In this section, we have developed a method called ISGE (**I**dentify, **S**elect and **GE**nerate) for optimising schemes in particular contexts. This method creates conceptual views from the CSG that are tailored to the particular specifications of each use case under study. Our goal is that these conceptual views fulfil the requirements of any working domain while reducing the complexity and the amount of time required to understand and adopt them.

The correct genomics data management, the efficient knowledge sharing, and the deep domain understanding we want to achieve strongly depend on the CSG. We envision our schema as the single point of truth from which a universal conceptualisation of the relevant genome properties common to any particular species is provided.

The CSG results from the research work presented in this Ph.D. thesis: it is more extensive, complete, and complex than its predecessors (i.e., the CSHG and the CSCG). The CSG excels at providing a comprehensive representation of the genome. However, efficiency, in terms of understandability and adoption, is crucial in real world use contexts. We can improve such an efficiency by reducing the size of the CSG, thereby delivering portions of the CSG that focus on the particularities of a given use case. In other words, why would we use the whole CSG if we are only interested in working with two of its views (e.g., the variation view and the pathway view).

We envision a more appropriate solution, namely, to provide a mechanism to hide those parts of the schema that are irrelevant for the specific use case under study, thus reducing noise and complexity. There is an important difference between the concepts of "view" and "conceptual view". On the one hand, a view of the CSG is a cluster of classes that share a common aspect. For instance, the variation view clusters those classes associated with changes in the DNA with respect to the sequence of reference. On the other hand, a conceptual view is what is generated when the ISGE method is applied, and it constitutes an independent schema itself (see Fig. 3.14 for a more visual clarification).

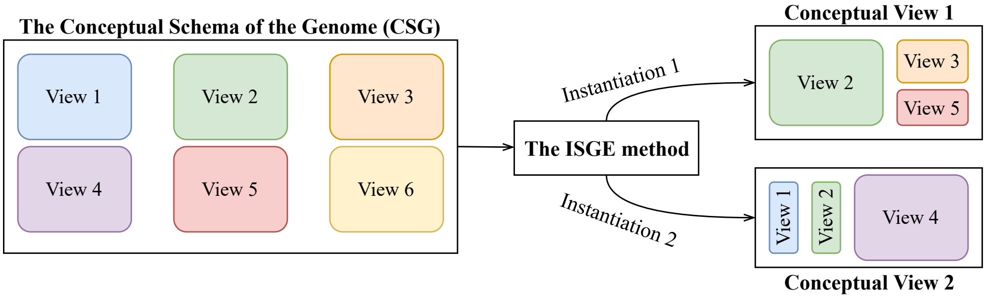

Figure 3.14: A visual clarification for the concepts of view and conceptual view. The CSG is divided into views, which cluster classes based on a common aspect. The ISGE method generates conceptual views, a subset of CSG classes tailored to a use case specification. For instance, View 1 of the CSG clusters 20 classes, none of which are part of Conceptual View 1, and only a tiny fraction is part of Conceptual View 2.

The ISGE method constitutes a framework that allows for the generation of conceptual views from the CSG that are adapted to the requirements of a specific use case. This framework facilitates using the CSG, reduces complexity, and increases the reuse of previously accumulated knowledge. The ISGE method minimises conceptual overload and allows researchers to focus on the relevant parts of the CSG.

3.4.2 Phases of the Method

The ISGE method is divided into three phases:

1. Identify: The relevant knowledge for the specific use case is identified in this phase. A set of interviews and focus groups are conducted with domain experts to make such knowledge as explicit as possible. After the initial discussions, a first draft of the use case description is generated. This draft allows for a better conceptualisation of the problem and aims to make implicit knowledge of the domain explicit. An additional benefit of Phase 1 is that it allows non-experts to clarify complex concepts, resulting in potential improvements to the model.

The artifact obtained from this phase is a detailed textual description of the use case under study. This artifact is used in the next phase to identify relevant classes that will be used to build the conceptual view.

2. Select: The selection of the required CSG classes is carried out in this phase. The textual description is mapped to the corresponding classes of the CSG, generating a first draft of the conceptual view. This correspondence complements the use case description, achieving full traceability from requirements to the conceptual view to be created. This process is carried out on a per-view basis. Table 3.3 shows the template used to carry out this process.

Sentence	**Class(es)**	**Explanation**
	On a per-view basis	
A sentence of interest that justifies including a set of classes from the CSG into the conceptual view.	the set of classes.	Any further clarifications regarding this sentence-classes association should be written here.

Table 3.3: This template associates portions of the artifact generated in the "identify" phase with classes of the CSG.

Once the first draft is generated, it is discussed and validated with domain experts. Two additional tasks emerge during these discussions. The first task is to evaluate the consistency of the model. For instance, this involve ensuring that no portions of the schema are isolated or it may involve dealing with gaps of knowledge. Imagine that a given domain contains data regarding genes and proteins but lacks data regarding transcripts. We must adapt to such situations and modify the schema by creating temporal links that collapse absent knowledge between classes. It is common for domain experts to identify such inconsistencies because they work with domain data and are aware of such limitations. Table 3.4 shows the template used to create such temporal links.

Original 3–tuples	**New 3–tuple**	**Explanation**
A set of 3–tuples in the form of "source class - relationship - destination class" that cannot be represented in the conceptual view	The new 3–tuple (using the same format) to represent the original 3-tuple(s)	The reasons that justify such change in the conceptual view.

Table 3.4: This template reports on the changes conducted to solve consistency issues (i.e., the creation of temporal links that did not exist in the CSG).

The second task is the identification of proposals to improve or enlarge the CSG. As the advances in genome knowledge and understanding are made, there may be instances where particularities of a use case cannot be appropriately described with the CSG. This situation is an excellent opportunity to make implicit (or even previously unknown) knowledge explicit with the ontological support of the domain experts working with such concepts. Appendix 3.5 shows the template used to propose improvements to the CSG.

Conceptual view	**Proposal**	
	Title	**Description**
The conceptual view from which the proposal has been generated.	Title of the proposal.	A more detailed proposal.

Table 3.5: This template describes proposals to improve or enlarge the CSG as a result of applying the ISGE method to a particular domain.

The artifacts obtained from this phase are the following: first, a document containing the mapping between the textual description and the classes of the CSG and, optionally, a document identifying model inconsistencies and how they were solved; second, a document with the set of identified proposals for improving the CSG.

3. Generate: The conceptual view is generated in this phase. The artifact obtained from this phase is a novel conceptual view that is tailored to the specifications of a given use case, with full traceability from requirements to model.

3.4.3 Conclusions

Throughout this section, we answered **RQ6**: Why/What/How to create a method to generate subschemes of the Conceptual Schema of the Genome? The ISGE method is a methodological framework that allows for creating conceptual views of the CSG. This method is the fifth contribution of this thesis.

Another benefit of this method is the flexibility we gain when facing different problems. The whole CSG can be used if the main goal is domain understanding or communication. However, suppose that the main goal is to improve genomics data management in a specific use case that only considers a small fraction of the CSG. In that instance, the ISGE method can create a conceptual view that is more appropriate for working with such data.

3.5 The Ontological Unpacking

So far, we have applied traditional conceptual modelling techniques. Now we aim to *use an ontology-based conceptual approach* to capture and better represent the particularities of genomics. (**G2.5**).

3.5.1 Introduction

The "ontological unpacking" process consists of a model-to-model transformation where the input is a traditional conceptual model. After this, a foundational ontology is used to generate an ontology-based conceptual model (see Fig. 3.15).

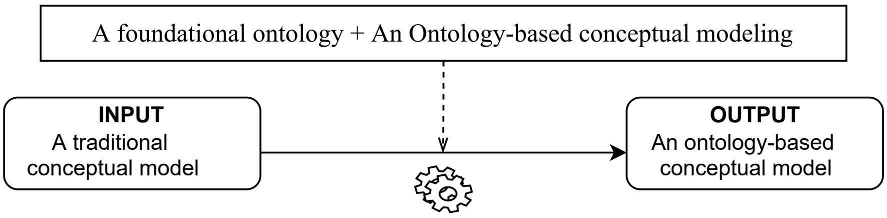
A foundational ontology + An Ontology-based conceptual modeling
INPUT
A traditional conceptual model
OUTPUT
An ontology-based conceptual model

Figure 3.15: An schematic representation of the Ontological unpacking process

We used OntoUML[12] to include the semantics defined by UFO. As mentioned above, this unpacking consists of transforming a "flat" UML Class Diagram model into a OntoUML model.

The ontological unpacking process has been carried out over a subset of classes of the pathway view of the CSG along with some additional classes from other views (for a complete representation of the considered schema, see `12-experiment_uml.pdf` file[13] in [180]). This is because the pathway view reflects very critical aspects of the genomics domain, including how genome elements interact over time to produce biological behaviour.

[12] OntoUML is a UML extension based on UFO to conduct ontology-driven conceptual modeling.
[13] `https://zenodo.org/record/7071090/files/12-experiment_uml.pdf`

We transformed our classic UML model into its ontologically well-grounded counterpart using OntoUML, an ontology-driven conceptual modelling language based on UFO. OntoUML provides us with stereotypes to characterise classes and relationships with the UFO constructs. Below, we present to the reader the fundamentals of UFO and the class stereotypes used during the ontological unpacking exercise.

UFO differentiates between Types and Individuals. Types are defined as abstract things we create to classify the world around us, whereas Individuals are particular instances of a given type. This relationship between an Individual and its Type is called instantiation. Additionally, there are Types that are first-order, meaning that they are instantiated by individuals, and Types that are higher-order, which are instantiated by other Types. Let us illustrate this with two examples:

- "Person" is a first-order Type, and the readers of this thesis are Individuals that instantiate the person Type.
- "Wolf" is a first-order Type that instantiates the "mammal" higher-order Type.

There are five principles that need to be considered to understand UFO:

- **Time**: UFO distinguishes between endurant and perdurant Individuals. Endurants are defined as objects that exist in time, including their existentially dependent properties. Perdurants are defined as events that happen within a specific time frame. For instance, this thesis is an Endurant, while the act of reading it is a Perdurant. Endurants are classified based on multiple dimensions, which are explained below.
- **Sortality**: It differentiates between Types that provide a uniform principle of identity to their instances (sortal) and Types that do not provide it (non-sortal). A principle of identity is what makes things what they are and allows us to distinguish between two Individuals unequivocally. An identity principle has a set of constraints regarding its nature: it points to a single Individual and to the Individual as a whole, and it always points to the same Individual. To illustrate, what is the identity principle for a person? Its name is not valid because it is not unique; its social security number is not valid because it is not universal. It could be the case that the answer to this question is the person's DNA sequence. As we can see, defining an identity principle is challenging. However, we know that it exists.

- **Rigidity**: It is a meta-property of Types associated with the instantiation of Types by Individuals. A Type can be rigid, non-rigid, or anti-rigid. It is rigid if it is instantiated in all possible scenarios in which its associated Individual exists. It is non-rigid if there is at least one scenario in which it is not instantiated, and its associated Individual still exists. It is anti-rigid if, in every possible scenario, it can cease to be instantiated and its associated Individual still exists. For instance, the author of this thesis is a Person in all possible scenarios in which he exists. This means that the Type person is rigid. However, he was a child during a past period of time, but not anymore. This means that the Type child is anti-rigid.
- **Relational dependence**: It indicates whether the classification conditions are based on a relational property or not. For instance, a marriage is externally dependent on the two people that got married.
- **Unity**: It is a Type's characteristic associated with how the parts contribute to the whole in a Type. There are three principles of unity: functional complex (the parts play different functional roles with respect to the whole), collectives (the whole has a uniform mereological structure), and quantities (amounts of matter).

Table 3.6 describes the Endurant stereotypes that we have used in the ontological unpacking exercise, indicating its sortality, rigidity, relational dependence, and unity.

Stereotype	Description	S	R	RD	PoU
Type	High-order Type whose instances are themselves Types.	Non-sortal	Rigid	No	—
Category	Cluster of properties that are shared by multiple identity providers.	Non-sortal	Rigid	No	—
Kind	An Identity provider Type whose part contributes in different ways to the whole.	Sortal	Rigid	No	Functional complex
Collective	An Identity provider Type whose parts are perceived in the same way by the whole.	Sortal	Rigid	No	Collective
Subkind	Rigid specialisations of rigid sortals.	Non-sortal	Rigid	No	—

Table 3.6 continues on the next page

Table 3.6 continued from previous page

Stereotype	Description	S	R	RD	PoU
Role	Specialisation of an identity provider instantiated in relational contexts.	Non-sortal	Anti-rigid	No	—
RoleMixin	Analogous to Role but for non-sortals.	Non-sortal	Anti-rigid	No	—
Historical-Role	Specialisation of an identity provider instantiated by its participation in an Event.	Non-sortal	Anti-rigid	No	—
Historical-RoleMixin	Analogous to HistoricalRole but for non-sortals.	Non-sortal	Anti-rigid	No	—
Phase	Specialisation of an identity provider instantiated by changes in intrinsic properties.	Non-sortal	Anti-rigid	No	—
PhaseMixin	Analogous to Phase but for non-sortals.	Non-sortal	Anti-rigid	No	—
Relator	Represents things that must exist in order for two or more Individuals to be connected. They depend on Individuals to exist.	Sortal	Rigid	Yes	Functional complex
Quality	A particular type of intrinsic property whose value is structured.	Sortal	Rigid	No	Functional complex

Table 3.6: The list of OntoUML stereotypes used in the ontological unpacking process reported in this thesis. The third column refers to the sortality principle; the fourth column refers to the rigidity principle; the fifth column refer to the relational dependence principle; the sixth column refers to the principle of unity

Table 3.7 describes the relationship stereotypes that we have used in the ontological unpacking exercise. The stereotypes described below are defined between Types; however, there are reflected at the Instance level (i.e., on Individuals).

Stereotype	**Description**
characterisation	It connects a Quality to its corresponding bearers.
creation	It connects an Endurant with the Perdurant in which it was created.
externalDependence	Given two events: A and B; A is externally dependent of B if A is existentially dependent on B and B is mereologically disjoint from A (i.e A is not part of B, B is not part of A, and A and B do not share common parts).
historicalDependence	Every event has a begin-point and end-point. This stereotype allows for describing temporal precedence of Events.
instantiation	It connects a Type with the stereotypes instantiating it.
material	It is derived from a Relator and constitutes a direct association between the Individuals connected by a Relator.
mediation	It connects a Relator and the Individuals it connects.
memberOf	It is the parthood relationship between a Collective and its parts
participation	It connects an Endurant with a Event in which it participated.
participational	Complex events can be composed of other events. This stereotype allows for defining the "has-part" relationship between events.
termination	It connects an Endurant with the Event in which it was terminated.

Table 3.7: The list of OntoUML relationship stereotypes used in the ontological unpacking process reported in this thesis.

We have carried out the ontological unpacking process to the pathway view of the CSG (see Fig. 3.16). The resulting OntoUML model can be seen in `13-experiment_ontouml.pdf` file[14] in [180]. The resulting conceptual schema has a sound and precise ontological commitment. The two top-level classes, Entities and Events, were represented as "plain" classes in the original UML schema. In the OntoUML schema, their conceptual characterisation is more precise thanks to the finer-grained constructs that differentiate between endurants and perdurants. The results are divided based on the changes associated with endurants (i.e., things that exist), perdurants (i.e., things that happen), and the participation of endurants in perdurants.

[14] `https://zenodo.org/record/7071090/files/13-experiment_ontouml.pdf`

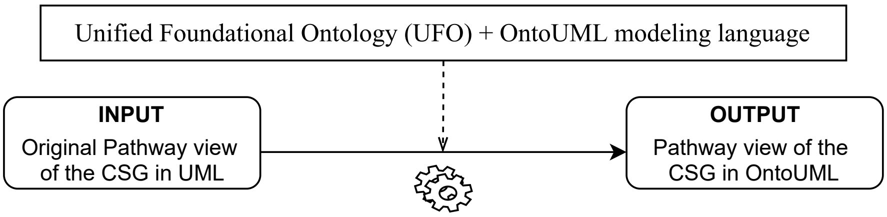
Unified Foundational Ontology (UFO) + OntoUML modeling language
INPUT
Original Pathway view of the CSG in UML
OUTPUT
Pathway view of the CSG in OntoUML

Figure 3.16: A schematic representation of how we carrried out the Ontological unpacking process using the pathway view of the CSG as input, UFO as foundational ontology, and OntoUML as modeling language

3.5.2 Endurants

The Entity class characterises a wide range of molecules with different identity principles; thus, we annotated this class with the «category» stereotype.

The first type of Entity is the Complex, which we stereotyped as a «kind». This type of entity is created when at least two biological Entities are combined. Each of these Entities, known as Component, contributes to the whole with a specific role. Although a Component is part of a whole, it does not lose its identity principle, and it can be detached from the Complex without being destroyed. We captured such particularities by annotating the Component with the «roleMixin» stereotype. This means that it is instantiated in relational contexts (i.e., when multiple Entities combine to form a Complex).

A «material» relationship between the Complex and the Component characterises how the parts are connected to the whole. We materialised such a relationship by means of the new ComponentInComplex class, stereotyped as a «relator». Since relators are the truth-makers of a relationship, ComponentInComplex captures the *interaction_type* and *stoichiometry* of the Component. An instance of ComponentInComplex is needed to connect a Component and a Complex. Since every Complex is composed of at least two Components, at least two instances of ComponentInComplex per Complex must exist.

The second type of Entity is Simple, which is stereotyped as «category». This class specialises into Polymer and Monomer, which is also annotated with the «category» stereotype because they do not provide an identity principle. The three types of Polymers, namely, NucleicAcid, BasicPolymer, and Peptide are stereotyped as «collective» because their internal structure is homogeneous (i.e., a chain of MONOMERS that contribute in the same way to the whole).

Finally, the ENTITYCLASS has been stereotyped as «collective» because, although they can group very diverse ENTITIES with different identify principles, they play the same role with respect to a specific PROCESS.

3.5.3 Perdurants

The EVENT class (renamed to BIOLOGICALEVENT) has been annotated with the «event» stereotype to represent that they are entities that happen over time, accumulating temporal parts. This also allowed us to extend the definition of events with the *start* and *end* attributes to support reasoning with Allen's time interval relations [181]. In the UML schema, we have a reflexive relationship connecting the EVENT class with itself. This relationship is ambiguous because it does not indicate whether it represents mere temporal precedence between two EVENTS or a strong causal connection. This ambiguity is solved in the OntoUML model by means of the «historicalDependence» stereotype, which makes explicit that if an EVENT of type A is historically dependent on an EVENT of type B, instances of A are necessarily preceded by instances of B, but not vice versa.

The characterisation of EVENTS is a direct instantiation of UFO's structural partonomy pattern [182]: there are two types of EVENTS, namely, atomic (i.e., PROCESS) and complex (i.e., PATHWAY), connected by an aggregation relationship. This relationship follows the weak supplementation axiom [182], which imposes that complex entities must be composed of at least two disjoint parts.

3.5.4 Endurants participating in Perdurants

The participation dimension represents ENTITIES' role in PROCESSES, and the OntoUML model expand the previous characterisation. First, we created a new class called PARTICIPATIONINBIOLOGICALEVENT stereotyped as «event». This class islocated between the TAKESPART and PROCESS classes and allows for a PROCESS to be explicitly divided into the individual participation of ENTITIES. Every instance of this new class is derived from parthood and existential dependence, and it is bound to a specific specialisation of TAKESPART (i.e., INPUT, OUTPUT, or REGULATOR, among other roles that may be discovered in the future). In the OntoUML model, events can be divided based on the atomic parts that compose them or on the individual participation of

Endurants. For instance, protein synthesis can be decomposed into atomic steps (i.e., initiation, elongation, and termination), creating segments that use temporal schemes as external references. It can also be decomposed into portions that capture the individual participation of molecules during the whole process (e.g., the participation of the ribosome, the participation of the mRNA strand, etc.).

Second, we annotated the TakesPart class and its specialised classes with the «historicalRoleMixin» stereotype. This change imposed a minimum cardinality of one in the association between the TakesPart class and the Process because, for an Entity to play the historical role, it must have mandatorily participated in an event.

Another improvement of the new schema is that we can explicitly represent the creation and termination of Entities. We created two new classes (i.e., ActiveEntity and DegradedEntity) and annotated them with the «phaseMixin» stereotype. This stereotype represents changes in the intrinsic properties of Individuals (in this case, whether or not it is destroyed). Representing the creation and termination of Entities was a missing feature that the ontological unpacking exercise revealed, thereby allowing us to characterise it. We also created the *creation_ date* attribute in Entity to identify when it was created.

3.5.5 Discussion

The resulting schema led to improvements in the representation of the genomics domain, including the characterisation of biological entities, the changes in biological entities over time, and the representation of chemical compounds.

Characterisation of biological entities

The "flat" semantics of UML does not consider the identity and rigidity dimensions, meaning that Types with the identity principle (e.g., «kind») are represented in the same way as those without it (e.g., «category»). Similarly, there is no distinction between rigid (e.g., «kind») and anti-rigid (e.g., «role») types. The ontological unpacking shows how not considering such dimensions can affect conceptual clarity. For instance, in the OntoUML schema, it is clear that a Protein i) is a type of Peptide, which is provides an identity principle, ii) has an homogeneous internal structure, and iii) shares a cluster of properties with DNA, although they share different identity principles.

Changes in biological entities over time

The «phase» stereotype enriches the representation of effects caused by events. In the OntoUML version, there is an additional dimension to describe whether an entity has been created or degraded. This allows for describing not only the cases considered in the UML schema (i.e., input, output, and regulator), but also those cases in which:

- An entity that is degraded as a result of a process.
- An entity that is created as as a result of a process.
- An entity that is degraded as a result of regulating a process.

Modelling changes in the intrinsic properties of entities (i.e., if it is degraded or not) was not possible in the UML version. The OntoUML version explicitly states that the creation and destruction of entities result from processes. Additionally, the new classes stereotyped as phases provides additional mechanisms for data correctness of the model. For instance, we can easily specify that ENZYMES that catalyse a process cannot be destroyed as a result of such a regulation (i.e., an ENZYME cannot instantiate the DEGRADEDENTITY «phaseMixin» and the CATALYSIS «historicalRoleMixin» in the same PROCESS). Specifying such constraints is a very appropriate way of preventing errors when instantiating and populating the model. However, they are difficult to consider in the UML schema.

The representation of chemical compounds

Thousands of chemical compounds exist in the human body. They interact with our proteins in a myriad of never-ending processes. In UML, they were represented with the BASIC CLASS, a type of SIMPLE entity. However, this characterisation was not clear enough to address questions such as: Can a chemical compound be a polymer? What is the monomer of a chemical compound that is a polymer? The UML schema could not differentiate appropriately between chemical compounds that are not polymers nor monomers, chemical compounds that are polymers, and the monomers of these polymers.

We created two new classes. The first, BASICPOLYMER, is stereotyped as a «collective» and represents chemical compounds that are polymers. The second, BASICMONOMER, is stereotyped as a «kind» and represents those chemical compounds that are monomers. These two classes, along with the old BASIC class that represents chemical compounds that are not polymers nor monomers, are clarified the schema.

3.5.6 Conclusions

Throughout this section, we answered **RQ7**: How to conduct ontology-driven conceptual modelling in genomics? The ontological unpacking process instantiated above consisted of a UML-to-OntoUML model transformation that allowed us to obtain an ontology-driven conceptual schema. This OntoUML schema is the sixth contribution of this thesis. UFO and OntoUML made it easier for us to capture some of the particularities of genomics better than the UML schema did, resulting in a more accurate and explicit schema.

3.6 Conclusions

Throughout this chapter, we have tackled **G2**: generate conceptual modeling artifacts to improve genomics data management. We answered the four research questions associated with G2.

First, we updated the Conceptual Schema of the Human Genome generating a new version of the CSHG, which has resulted in a more complete, detailed, and generic schema, easing how domain experts extract knowledge and how non-experts are introduced in the domain. This update answers **RQ3** (associated with **G2.1**): Why/What/How to extend and update the CSHG?

After updating this schema, we explored the conceptualisation of the genome for a species other than humans. Together with the IVIA research group, we studied how they work with genomics data from the agrifood domain, which resulted to the Conceptual Schema of the Citrus Genome. Creating the CSHG answered **RQ4** (associated with **G2.2**): Why/What/How to generate a conceptual schema of the genome for non-human species?

Both conceptual schemes are different instances of the same conceptualisation process: the genome. As a result, we answered **RQ5** (associated with **G2.3**): Why/What/How to generate a conceptual schema of the genome that is species-independent? As a result, we generated the CSG and updated it based on the feedback from domain experts. After answering these three research questions, our conceptual schema is ready to be used in real-world scenarios for the first time.

The CSG is a comprehensive representation of the genome intended to improve domain understanding, knowledge sharing, and communication between experts from different domains (e.g., computer scientists and geneticists). To ease the use of the CSG and facilitate its adoption in real-world use cases, we developed the ISGE method. This method generates narrowed-down conceptual schemes that allow more efficient data management and integration processes. The generation of this method allowed us to answer **RQ6** (associated with **G2.4**): Why/What/How to create a method to generate subschemes of the Conceptual Schema of the Genome?

Finally, we performed a UML-to-OntoUML model transformation (i.e., an ontological unpacking) to generate an ontology-based conceptual schema grounded on UFO. This OntoUML schema allowed us to answer **RQ7** (associated with **G2.5**): How to conduct ontology-driven conceptual modelling in genomics? Another benefit is that this OntoUML schema has support for formal verification, validation, and reasoning by automatically generating OWL specification for the model.

Six contributions have been presented throughout this chapter. The first three contributions (i.e., Version 3 of the CSHG, the CSCG, and Version 1 of the CSG) constituted intermediate steps to achieving the final three contributions (i.e., i.e., Version 2 of the CSG, the ISGE method, and the OntoUML schema) aimed at improving the existing problems in genomics data management.

Chapter 4

Treatment Validation

In this Chapter, we confirm and validate our contributions and research results (**G3**). To this aim, we answer the following research questions:

- To what extent are the contributions of this thesis useful in a human genomics context? (**RQ8**).
- To what extent are the contributions of this thesis useful in an agri-food genomics context? (**RQ9**).
- Does ontology-driven conceptual modelling capture domain particularities better than traditional conceptual modelling? (**RQ10**).

To answer **RQ8**, we developed a conceptual model-based platform called The Delfos Oracle that allows for the identification of relevant variations associated with different diseases. A lab validation was carried out to validate the Delfos Oracle. This validation comprised two phases. First, we conducted two empirical experiments to identify relevant DNA variations associated with Alzheimer's Disease and muscular dystrophies. Second, we carried out a semi-unsupervised experiment where several individuals performed a set of tasks using the platform.

To answer **RQ9**, another conceptual model-based platform called CitrusGenome was developed. This second platform implements an automated workflow for performing efficient comparative genomic studies over dozens of DNA sequences of citrus crops. To validate the tool, we asked domain experts to conduct a set of supervised comparative analyses. Participants were asked to think aloud to formulate their first impressions when using CitrusGenome. Afterwards, we carried out a focus group where domain experts shared their thoughts and discussed the utility of the tool. Finally, an unsupervised comparative study was conducted whereby participants were asked to identify variations associated with a specific trait. To test for validity, the results of this study were compared with the existing body of knowledge in order to test their validity.

To answer **RQ10**, an empirical study was conducted to compare the generated OntoUML schema to its corresponding UML counterpart. We measured effectiveness, efficiency, and satisfaction. The experiment was conducted by 20 subjects who were students from two computer science classes in their fourth year.

The chapter is structured as follows:

Section 4.1 – Validates the CSG and the ISGE method in a human genomics context.

Section 4.2 – Validates the CSG and the ISGE method in an agri-food genomics context.

Section 4.3 – Assesses the utility of the ontological unpacking.

Section 4.4 – Reports conclusions.

4.1 The Delfos Oracle Platform

We start this section with by briefly introducing the newly developed platform and the problems to be solved. Then, we report on the use of the ISGE method to generate the use case-specific conceptual view and introduce the platform. Finally, the validation is carried out.

4.1.1 Introduction

The first platform that validates the contribution of this thesis is a Genome Information System called the Delfos Oracle Platform. This platform allows for extracting, integrating, prioritizing, storing, and visualizing clinical genomics data associated with diseases (i.e., DNA variations that cause genetic disorders and a plethora of complementing information supporting such associations).

One of the many consequences of the problems associated with genomics data management is that personalised treatments in precision medicine are not as personalised and accurate as they should be. In other words, knowledge generation processes are limited, and the quality of their results is questionable.

The knowledge needed to provide correct personalised treatment is spread across heterogeneous databases, isolated from each other, and often with insufficient data quality controls. This situation causes interoperability and data quality problems. However, data correctness not only depends on data quality controls but also on errors that may appear due to other reasons, such as failures during data collection and filtering processes or experiments carried out over non-representative population samples. [183]. Adding to this, even if domain experts succeed in integrating the data they need, they encounter the challenge of extracting knowledge from a massive volume of data, of which only a tiny portion is relevant. As a consequence, identifying DNA variations responsible for causing disorders with a high-quality level of evidence is a titanic task.

The Delfos Oracle platform aims to improve the identification of relevant and high-quality variations. To do so, it implements the four phases of the SILE method, which provides a systematic approach to manage genomics data. First, the Search phase determines the most suitable data sources and integrates their data; Second, the Identification phase identifies, filters and classifies the relevant data; Third, the Load phase stores the filtered data for further analysis and exploitation; Fourth, the Exploitation phase eases knowledge generation from the information obtained in the three previous stages of the method. More information can be found in [184].

The Delfos Oracle is composed of four modules, one per each phase of the SILE method. The first module, called Hermes, searches for reliable data sources and integrates their data. The second module, called Ulises, identifies those pieces of data that are relevant by means of an AI algorithm. The third module, called Delfos, is in charge of storing relevant data adequately. The fourth module, called Sibila, exploits the data to ease knowledge generation processes. Sibila aims to cover a critical dimension that is often forgotten in genomics: user interface design. The use of sound conceptual modeling techniques can be a helpful tool for designing and developing user interfaces that are intuitive and easy to use [42].

This platform is conceptual model-based, meaning that a conceptual model has guided the process:

- **The ISGE method** has been used to generate a conceptual view tailored to the specific needs of the platform.
- **The platform architecture** has been designed and implemented using the generated conceptual view:
  - For Hermes, the model has served as the template for the data-transformation processes needed to integrate the data from the different data sources.
  - For Ulises, the model has helped in mapping the algorithm criteria to the data attributes.
  - For Delfos, the model has driven the design of the database's physical schema and the defined quality checks
  - For Sibila, the model has guided the design and implementation of the User Interfaces.

4.1.2 The ISGE method

The Delfos Oracle Platform core is the Conceptual Schema of the Genome for Delfos (CSGD) (see `01-csgd.pdf` file[1] in [180]), a conceptual view generated from the CSG using the ISGE method (see Fig. 4.1). This conceptual view provides the conceptual structure that is needed to efficiently connect all of the data sources under a holistic perspective. The CSGD focuses on variations and their associations with diseases without considering other specific parts of the Conceptual Schema of the Genome, such as the structure of the proteins or the DNA.

[1] `https://zenodo.org/record/7071090/files/01-csgd.pdf`

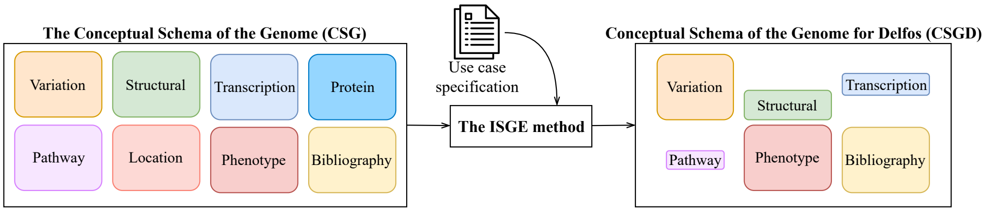

Figure 4.1: Visual representation of the instantiation of the ISGE method for the human case.

1. Identify

The main goal of the Delfos Oracle is to integrate genomics data from several data sources associated with diseases, integrating them and calculating the clinical actionability of DNA variations. Since genomics data changes continuously, it is crucial to track the different data sources and how they evolve. Another dimension of such temporal variability is the coexistence of several assemblies in which variations are identified and reported.

For each disease, we are interested in the number of chromosomes and genes affected, the number of variations associated with the phenotype, and the specific sources from which phenotype data has been retrieved.

For each variation, we are interested in its general information (i.e., its name, type, alleles, location, and affected gene) and the clinical significance and clinical actionability of every associated phenotype. This significance and actionability are calculated based on the information provided by submitters, scientific bibliography, and studies with statistical associations. Also, we are interested in the HGVS expressions that allow for identifying such variations. Finally, we need to know the data sources from which the variation information has been retrieved.

The Delfos Oracle mainly focuses on mendelian disorders, such as cystic fibrosis, Duchenne muscular dystrophy, and Alzheimer's disease when affecting young people (i.e., the early onset form of the disease). Mendelian disorders are caused by variations produced on single genes.

2. Select

The selection of the relevant pieces of information starts by mapping the textual description obtained in the first phase. The location and protein views are not considered because they are irrelevant with respect to the provided description. Concerning the rest of the CSG views:

Sentence	**Class(es)**	**Explanation**
Bibliography and DataBank View		
The main goal of the Delfos Oracle is to integrate genomics data from several data sources associated with [...] For each disease, [...] specific data sources [...] For each variation, [...] the data sources [...]	Databank, DataBankElement	These classes are used to represent the different data sources from which variations, diseases, and other relevant genomics data is obtained.
Since genomics data changes continuously, tracking the different data sources and how they evolve over time is crucial.	DataBankVersion	This class is used to track how genomics data changes over time when external sources are updated.
This significance and actionability are calculated based on the information provided by submitters, [...]	Submission, Submitter,	The classes used to represent information provided by submitters.
This significance and actionability are calculated based on the information provided by [...], scientific bibliography, [...]	BibliographyReference	This class models the concept of scientific bibliography.
This significance and actionability are calculated based on the information provided by [...] studies with statistical associations.	Study, StatisticalAssociation, Population, AlleleFreq, GenotypeFreq	These are the classes required to appropriately characterize studies with statistical associations.

Table 4.1 continues on the next page

Table 4.1 continued from previous page

Sentence	Class(es)	Explanation
Location		
No relevant classes have been identified for this view.		
pathway		
For each variation, we are interested in [...] location, and affected gene.	Entity, Simple, Polymer, NucleicAcid	We incorporated the hierarchical definition of chromosomes and genes. The chromosome is required for identifying the location of a variation. The gene is required for identifying the genes affected by a variation.
Phenotype		
The main goal of the Delfos Oracle is to integrate genomics data from several data sources associated with diseases, [...]	Phenotype, Disease	Since we search for variations associated with diseases, we require these two classes to represent them.
[...] integrating them and calculating the clinical actionability of DNA variations. For each variation[...] the clinical significance and clinical actionability of every associated phenotype.	Certainty, Actionability	The significance and actionability concepts are captured by their corresponding classes.
Protein		
No relevant classes have been identified for this view.		
Structural		
[...] the coexistence of several assemblies [...] For each variation, we are interested in its [...] affected gene [...]	Assembly	We use this class to identify the working assemblies. Variations are located with respect to a specific assembly.

Table 4.1 continues on the next page

Table 4.1 continued from previous page

Sentence	**Class(es)**	**Explanation**
For each variation, we are interested in [...] location [...]	Chromosome, Sequence, Position	The chromosome is required for identifying the location of a variation. The VariationPosition class specialises from the Position class; thus, it is needed.
For each variation, we are interested in its [...] affected gene [...]	ChromosomeElement, ChromosomeElementPosition	To identify the genes affected by variations, we need to be able to locate them, which is accomplished in the CSG with these classes.
	Transcription	
For each disease, we are interested in the [...] genes affected [...] – For each variation, we are interested in its [...] affected gene	TranscriptableElement, Gene	These classes represent the concept of gene.
	Variation	
[...] integrating them and calculating the clinical actionability of DNA variations. For each variation [...]	Variation, Imprecise, Precise, Indel, Inversion, Deletion, Insertion	We use all of these classes to model the concept of variation.
[...] interested in the HGVS expressions that allow for identifying such variations.	HGVSExpression	This class represents the HGVS expressions associated with variations.

Table 4.1: Mapping carried out during Phase 2 (Select) of the ISGE method for generating the CSGD

We evaluated the classes selected to create the conceptual views and did not identify any inconsistencies. Nor did we identify any proposals for improving or enlarging the CSG.

3. Generate

The CSDG is composed of 29 classes (see Table 4.2); variation, phenotype and bibliography views are the most important. The views considered irrelevant, such as the protein view or the location view, have been removed from the final CSDG. This conceptual view can be seen in (see `01-csgd.pdf` file[2] in [180]).

View	**Concepts**		**Completeness**
	CSG	**CSGD**	
Bibliography	11	11	100%
Location	10	0	0%
Pathway	33	5	15.15%
Phenotype	4	4	100%
Protein	19	0	0%
Structural	11	6	55%
Transcription	17	2	11.76%
Variation	11	9	81.82%
TOTAL	115	29	25.21%

Table 4.2: Number of CSG concepts considered to build the CSGD view, which supports the Delfos Oracle platform.

4.1.3 The Platform Architecture

The four modules that compose the platform are the following:

- Hermes – This module comprises a set of libraries written in R that, guided by the CSGD, extracts and integrates data from the following data sources: ClinVar, Ensembl, GWAS catalog, and LOVD.
- Ulises – This module is a web-based automated workflow that classifies the variations previously integrated by Hermes. It is implemented using JavaScript and the node.js and react libraries. This module allows for rapid filtering of data. Ulises applies a rule-based algorithm to each variation, classifying them based on their associated evidence. As a result, every variation is accepted or rejected. Accepted variations are classified as either *to follow up*, *limited evidence*, *moderate evidence*, or *strong evidence*.

[2] `https://zenodo.org/record/7071090/files/01-csgd.pdf`

- Delfos – This module is a database and a web-based tool to load the data retrieved by Ulises. It is implemented using MySQL and Javascript with the node.js and react libraries. This module constitutes the basis for knowledge-generation processes. Added to this, Delfos performs a quality check prior to loading the data to ensure that the required minimum data quality standards are met.
- Sibila – This module is a web-based platform that allows for identifying relevant variations. It is implemented using JavaSript and the node.js and react libraries. These variations can be filtered based on the genes and phenotypes they are associated with.

Let us illustrate the benefits of following a conceptual model-based approach in terms of User Interface (UI) design using Sibila's UI that displays the data associated with a specific variation as an example. These details include the specific characteristics of a variation, the associated phenotypes, and the external sources where the variation can be found.

Table 4.3 shows the visualisation patterns we selected for designing this UI. These patterns are associated with the pieces of the CSG (i.e., the relevant classes and their attributes) required to instantiate it correctly. For instance, pattern 1 displays the *name* and *date* attributes of the VARIATION class. For this specific UI, we selected four patterns (i.e., the Card, Module Tabs, Chunking, and Tagging) that were instantiated ten times.

Applying the ISGE method guarantees that the transformation of the requirements into the UI is clear, because a conceptual schema explicitly supports it. For instance, according to the requirements gathered in the Identification phase, variations must display the bibliography associated with their clinical significance. This requirement is implemented using pattern 9, which requires the *title*, *authors*, and *date* attributes from the BIBLIOGRAPHY class instances of interest and the *url* attribute from the DATABANKELEMENT class instances of interest. These two classes were selected in the Selection phase of ISGE when mapping this requirement to the CSG (see Table 4.1).

ID	Pattern	Applied to	Class	Attribute(s)
1	Card	Displays the name and date of the variation	Variation*	name*, date*
2	Card	Displays the rest of the variation information	—	—

Table 4.3 continues on the next page

Table 4.3 continued from previous page

ID	Pattern	Applied to	Class	Attribute(s)
3	Module Tabs	Separates the data into sections that can be accessed using flat navigation	—	—
4	Chunking	Groups the general information of a variation	Variation*	type*
			Precise	ref*, alt*
			Chromosome[]*	name*
			Assembly[]	name*
			VariationPosition[]	start*, end*
			Gene[]	name*
5	Chunking	Groups the HGVS expressions of a variation	HGVSExpression[]	expression*
6	Tagging	Labels the clinical actionability and classification of a variation for a phenotype	Actionability*	clinical_actionability*, classification*
7	Chunking	Groups the information of a phenotype. It contains 6, 8, and 9 patterns	Phenotype[]*	name*
8	Chunking	Groups the significances of a variation for a phenotype	Certainty[]*	clinical_significance*, method*, criteria*
9	Chunking	Group the bibliography of a variation for a phenotype	Bibliography	title*, authors*, date*
			DataBankElement[]	url*

Table 4.3 continues on the next page

Table 4.3 continued from previous page

ID	Pattern	Applied to	Class	Attribute(s)
10	Chunking	Group the references of a variation in external data sources	DataBank[]*	name*
			DataBankElement[]*	url*

Table 4.3: List of selected patterns for a specific User Interface in the design of Sibila. The asterisk indicates mandatory data. Square brackets indicate an array of data.

After selecting the patterns and associating them to the corresponding classes and attributes of the conceptual view, we designed the UI. `00-sibila_implementation.png` file[3] in [180], which shows the final UI with the selected patterns implemented.

4.1.4 Validation of the Platform

Validating the tool allows us to demonstrate that our artifacts create solutions that mitigate the existing problems of genomics data management. Our goal with this validation is to determine whether the Delfos Oracle improves the identification of relevant and high-quality variations.

The validation comprised two phases. The first phase tested whether Delfos eased the identification of relevant DNA variations associated with a given disease. We conducted two supervised experiments where subjects tried identifying relevant DNA variations associated with two diseases using the platform. Second, we tested knowledge generation in terms of efficiency, effectiveness, and user satisfaction by conducting a semi-unsupervised experiment where several individuals performed a set of tasks using the platform.

First Phase

Two supervised experiments were conducted in a very restricted environment to answer the following questions:

- Does the Delfos Oracle improve the identification of relevant variations associated with Alzheimer's Disease?
- Does the Delfos Oracle improve the identification of relevant variations associated with muscular dystrophies?

[3] `https://zenodo.org/record/7071090/files//files/02-sibila_implementation.png`

Alzheimer's disease is the most common type of dementia and is characterized by cognitive impairment (i.e., trouble remembering, learning new things, concentrating, or making decisions that affect their everyday life). Although Alzheimer's Disease is more common among the elderly, Early Onset Alzheimer's Disease (EOAD) can also affect young people. Considering that EOAD is strongly associated with genetic causes [185], it is a good candidate for use with the Delfos Oracle platform. A master's thesis studied the quality of the data associated with EOAD in the Clinvar, Ensembl, and GWAS data repositories (i.e., 354 variations containing 398 variation-phenotype associations[4]) As a result, 351 variation-phenotype associations were rejected, 33 were accepted with limited evidence, 14 were accepted with moderate evidence, and no variations were accepted with strong evidence. [186].

Muscular dystrophies are a set of hereditary disorders linked to the X chromosome, and their main consequence is the progressive deterioration of muscular tissues [187]. Duchenne muscular dystrophy (DMD) is the largest human gene and it codes for several isoforms of dystrophin, an essential protein for connecting cytoskeleton and muscle fibers to the extracellular matrix. Becker Muscular Dystrophy (BMD) is similar to DMD, but its symptoms are less severe. When the functionality of the DMD gene occur, one of the muscular dystrophies above mentioned is manifested, depending on the specific alterations. A bachelor's thesis studied the data associated with two types of muscular dystrophy, namely, DMD and BMD [188]. The former is more severe than the latter. 2,561 variations for DMD and 122 variations for BMD obtained from the ClinVar, Ensembl, GWAS, and LOVD data repositories were analyzed. As a result, almost 70% of variations associated with DMD and 50% of variations associated with BMD were discarded, and none was accepted with strong evidence.

We successfully integrated data from different data sources, including Clinvar, Ensembl, GWAS, and LOVD. Besides, the number of variations identified as relevant decreased dramatically because they did not meet minimum data-quality requirements. None of the variations analysed in the Master's thesis and the Baherlor's thesis were accepted with strong evidence (i.e., those variations evaluated by an expert panel or following a clinical guideline). This result shows how the genomics data chaos complicates obtaining relevant data and reinforces the need for solutions for better genomics data management that improves genetic analyses. These results indicate that the Delfos Oracle allows for a more efficient identification of relevant variations (see Table 4.4).

[4] Some variations are associated to multiple subtypes of EOAD.

Phenotype	Variations	Accepted			To follow up	Rejected
		Limited	Moderate	Strong		
EOAD	398	33	14	0	1	350
DMD	2,581	757	43	0	15	1,766
BMD	122	40	7	0	14	61
TOTAL	3,101 100%	830 11.69%	64 2.06%	0 0%	30 0.97%	2,177 85.28%

Table 4.4: Summary of the variations analysed using the Delfos Oracle platform. We considered three diseases, namely, Early Onset Alzheimer's Disease, Duchenne Muscular Dystrophy, and Becker Muscular Dystrophy.

Second Phase

After the two supervised experiments reported above, we expanded the validation scope to focus on studying knowledge generation. Instead of supervising how the Delfos Oracle is used individually, we let several users work with the Delfos Oracle simultaneously iand semi-unsupervised. We conducted an empirical experiment that lasted for two sessions. A group of master's students used the Delfos Oracle to evaluate the process of identifying variations associated with a specific phenotype. Three dimensions were measured: effectiveness, efficiency, and user satisfaction.

Regarding effectiveness, we defined a set of metrics for each module of the Delfos Oracle:

1. The Hermes module: One metric to evaluate if subjects found the information in the proposed database and another metric to evaluate if subjects found information in additional databases. Both metrics have two possible values: 1 (correct) or 0 (incorrect).

2. The Ulises module: One metric to evaluate if subjects successfully applied the filters and another metric to evaluate if subjects correctly confirmed the results. Both metrics have two possible values: 1 (correct) or 0 (incorrect)

3. The Delfos module: One metric to evaluate if subjects loaded the data. This metric has two possible values: 1 (correct) or 0 (incorrect).

4. The Sibila module: One metric to evaluate the percentage of proposed queries answered and another metric to evaluate the percentage of queries

answered correctly. Both metrics are represented through a 0 to 100 percent.

The experiment results indicated that the effectiveness varied from one module to another. Although the effectiveness in the Hermes, Ulises, and Delfos modules was 100%, it decreased in the case of Sibila . Also, 22% of the students did not complete the empirical experiment's last step. The results of the students that did not complete the last step are associated with the typical risks of performing an empirical experiment across two sessions: the extra pressure associated with the limited time they had to complete the experiment influenced the results negatively. Additionally, some students did not have the minimum biological background required to answer the proposed questions correctly.

Regarding efficiency, we tracked the time necessary to perform each task:

- Hermes: the time needed to obtain data from both the proposed and additional databases.
- Ulises: the time required to filter the data.
- Delfos: the time required to load the data.
- Sibila: the time that users spent answering the proposed queries.

The most efficient modules were Ulises and Delfos (requiring less than 5 minutes), which was expected since they are the only modules that are fully automated. Obtaining the data with Hermes took most subjects between 90 and 270 minutes, and only one subject finished in 45 minutes; Answering the proposed queries with Sibila took less than 90 minutes for 80% of the subjects, while one subject required 180 minutes.

Regarding satisfaction, we used MAM questionnaires [189] based on the work of Moody et al. [190]. We measured Perceived Ease of Use (PEOU), Perceived Usefulness (PU), and Intention to Use (ITU) using 5-point Likert Scale questions. There is one MAM questionnaire per module, and the subjects completed them when they finished with each module.

The MAM questionnaires indicated that subjects considered the Delfos Oracle a valuable and easy-to-use platform. Approximately 60% of subjects opted for "agree" or "fully agree" for most questions of the Hermes MAM questionnaire; the results were better for PU, where this percentage rose to around 80%. The three remaining MAM questionnaires reported better results, and a very low percentage of subjects reported negative answers. All of the subjects managed to use the Delfos Oracle satisfactorily, even though they had no prior experience in genomics.

We have reported promising results for the three dimensions: users could identify high-quality variations associated with diseases in just two sessions, generating knowledge with a high percentage of effectiveness, efficiency, and satisfaction.

4.1.5 Conclusions

Throughout this section, we answered **RQ8**: To what extent are the contributions of this thesis useful in a human genomics context? We used the CSG and the ISGE method to generate a conceptual view that supported the generation of the Delfos Oracle. The Delfos Oracle showed that it is capable of integrating and filtering genomics data to improve the identification of high-quality variations. Conceptual model-based development of user interfaces led to efficient Human-Computer Interaction strategies that support effective data management.

Once we finished validating the platform in a lab environment (i.e., TRL–4[5]), the Delfos Oracle is ready to be tested in real-world scenarios. At the time of writing this thesis, the Delfos Oracle was being used in a research project to prioritise and filter variations identified in patients. This process will allow us to generate better diagnostic reports that can improve clinicians' decision-making. At the end of the project[6], we will be able to gather valuable feedback regarding the use of the Delfos Platform in a real-world, clinical context.

4.2 The CitrusGenome Platform

We start this section with a brief introduction to the newly developed platform and the problems to be solved. Then, we report on the use of the ISGE method to generate the use case-specific conceptual view and introduce the platform. Finally, the validation is carried out.

[5] Technology validated in lab, according to ISO 16290:2013.

[6] INNEST/2021/57 — Intelligent system for clinical decision support in precision medicine.

4.2.1 Introduction

The second platform that validates the contribution of this thesis is a Genome Information system, called CitrusGenome, which provides an efficient, user-friendly SNP discovery tool to perform genomic comparative studies. CitrusGenome is a conceptual model-based tool that operates over a database containing genomic variations (SNPs and INDELs) from 57 citrus genome sequences of the most relevant citrus species that have been annotated with additional data to provide a more holistic perspective.

Genome resources in citrus have increased over the last decade and, as a result, are currently reasonably abundant [191]. We collaborated with the *Instituto Valenciano de Investigaciones Agrarias* (IVIA) to improve how they perform comparative genomics analysis. Comparative genomics typically focuses on variations of gene content, transposable elements, large genome rearrangements, structural variations, and small polymorphisms. Among the latter, single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELS) are of great importance for plant breeding since they have proven to be major genetic determinants of relevant characteristics of agricultural interest, such as sweetness.

The number of available genomes has increased exponentially as the price of sequencing technologies has dropped. Because of this, IVIA has been able to generate a considerable amount of data from several citrus varieties. This situation offers new opportunities for SNP discovery, but it also poses several challenges related to computational efficiency, automation, data management expertise, and the inherent limitations of the several existing data types.

Integrating several types of genomic information, including gene topology and functionality characterisation, functional annotation, biochemical pathway information, and protein domain composition, is required prior to any analysis. Moreover, such analyses are particularly complex because citrus data poses peculiarities that do not occur in other areas of study in genomics. For instance:

- Clonal propagation and the occurrence of somatic mutations produce high rates of tissue chimerism, increasing the amount of background noise in SNP detection.
- Introgression events in domesticated varieties and genome rearrangements (deletions or translocations) create an unevenly distributed heterozygosity pattern across the genome.

The particularities reported above encourage using tools to visualise SNP distribution, which is essential for accurate interpretation.

Although Python scripting has been a feasible procedure for a limited number of these analyses for some of the IVIA researchers, SNP query at the whole-genome level involving several individuals is time-consuming. Also, in most cases, this scripting is limited to people with specific knowledge and computing skills. As a result, researchers without specific computing skills require a tool that:

1. Provides a scripting-free way to perform whole-genome queries, considering the many data types used in such queries.
2. Allows users to establish systematic and easy-to-apply strategies to reduce the amount of background noise when performing SNP detection.
3. Gives users a handy way of visualising and interacting with SNP distribution among chromosomes.

This platform is conceptual model-based, meaning that a conceptual model has guided the process:

- **The ISGE method** was used to generate a conceptual view tailored to the specific needs of the platform.
- Tha platform architecture was designed and implemented using the generated conceptual view:
    - For the database, the model was used to derive its physical schema.
    - For the backend, the model allowed for the precise identification of those attributes of interest that need to be queried to generate the desired results.
    - For the frontend, the model guided the design and implementation of the User Interfaces.

4.2.2 The ISGE method

The core of CitrusGenome is the Conceptual Schema of the Genome for Citrus (CSGC) (see `07-csgc.pdf` file[7] in [180]), a conceptual view generated from the CSG using the ISGE method (see Fig. 4.2). As with the human case, this conceptual view allows for the linking of data sources and formats with a holistic perspective. The CSGC focuses on variations and their consequences at the structural level, considering alterations in the transcription and other relevant biological pathways.

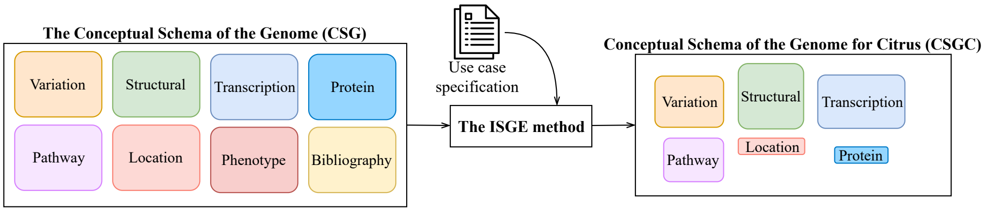

Figure 4.2: Instantiation of the ISGE method for the citrus case

1. Identify

Similar to the Delfos Oracle, we started by interviewing domain users and observing how they work, which allowed us to characterise them and define their specific tasks. Identifying and analysing the tasks involved in prioritising genetic variations allowed us to understand both the user mental model (i.e., how they think the variation prioritisation process works) and the particularities of the domain under study. The prioritisation of genetic variations (i.e., SNP and small INDEL variations) that might alter the expression on certain desired plant characteristics, involves complex analyses that are divided into three groups (for more information regarding the specification of these tasks and how they are represented in a CTT tree, refer to [192]):

1. Select Variety Groups: There are dozens of sequenced citrus varieties, and it is arduous to work with multiple vaarieties because of the large amount of data that is contained on each of them. In order to work with the varieties, bioinformaticians have to select and group them based on specific characteristics. Two groups are then created: one group contains varieties that highly express a characteristic of interest, and the other

[7] `https://zenodo.org/record/7071090/files/07-csgc.pdf`

group contains varieties that do not express the characteristic of interest. For instance, the sweetness of a specific variety of fruits.

2. Compare Groups: There are a plethora of variables to consider when filtering the data. Domain experts must reduce the existing amount of genomics data by applying several conditions, as a previous step, before comparing the groups of varieties. These conditions include data quality thresholds (i.e., the DP and GQ values of the variations identified in citrus), location in terms of scaffold and genomics regions (i.e., genes, promoters, introns, exons, mRNAs, coding regions, and untranslated regions), the predicted impact of variations, the gene product that is altered by the variation (i.e., proteins, enzymes, or their domains), the pathway where altered gene products participate, or the cellular component where altered gene products carry out their function. Also, researchers need a report on the applied filters in order to manage them. Considering the filter conditions, the groups of varieties have to be compared in order to extract their differences at a genotype level (i.e., genetic variations).

3. Visualise: The amount of data obtained after performing Task 2 can become unmanageable, and the bioinformaticians require fluidity to examine the gene products to enable the identification of potential genetic variations of interest. By "examine" we mean to i) show how the data are distributed based on specific criteria and ii) interact with the data by showing or hiding data columns and performing data.

Another relevant field of study for citrus is evolutionary genomics. The origin of citrus is still a matter of controversy, and the IVIA is currently generating knowledge in this area.

2. Select

We carefully evaluated the information gathered from domain experts and mapped it to the CSG. The bibliography and databank view was not considered because we focused on generating knowledge rather than integrating it. The phenotype view was not considered because there is no knowledge associated with the role of variations in specific phenotypes (see Table 4.5).

Sentence	**Class(es)**	**Explanation**
	Bibliography and DataBank View	
	No relevant classes have been identified for this view.	

Table 4.5 continues on the next page

Table 4.5 continued from previous page

Sentence	Class(es)	Explanation
	Location View	
[...] the cellular component where [...]	CellularComponent, Location	These classes represent the cellular locations where gene products carry out their activity.
	Pathway View	
[...] the pathways where altered gene products carry out their function [...]	Event, Pathway, Process	These classes represent the the set of pathways that occur in any living being.
[...] gene products carry out their function [...]	TakesPart, Input, Output, Regulator	These classes represent the functions of gene products with respect to the biological processes that compose pathways.
[...] location in terms of [...] genomics regions [...] the predicted impact of variations [...] [...] gene product that is altered [...]	Entity, Simple, Polymer, NucleicAcid, DNA, RNA	We incorporated the hierarchical definition of the genomics region analyzed, the biological elements considered for predicting the impact of variations, and the studied gene products.
	Phenotype	
	No relevant classes have been identified for this view.	
	Protein View	
the gene product [...] proteins	Isoform	The CSG models the proteins by means of their specific isoforms.
the gene product [...] enzymes	Enzyme	The class used to characterize enzymes in the CSG.
the gene product [...] domains	PrimaryStructureElement, Region, SecondaryStructureElement, TertiaryStructureElement, Domain	These are the minimum set of classes required to model represent the domains that compose proteins.
	Structural View	

Table 4.5 continues on the next page

Table 4.5 continued from previous page

Sentence	Class(es)	Explanation
There are dozens of sequenced citrus varieties [...]	Species, Individual	These classes allow for representing the sequenced citrus varieties.
[...] based on specific characteristics. The sweetness of [...]	SpeciesCharacteristic	This class is used to define the characteristics of sequenced citrus varieties.
location in terms of scaffold [...]	Scaffold, ChromosomeSequence	These classes are used to represent the scaffold concept and link it to the chromosome.
[...] genomics regions [...]	Chromosome, ChromosomeSequence, Position, ElementPosition	These classes are required to locate variations in specific genomics regions.
Transcription View		
location in terms of [...] genes [...]	TranscriptableElement, Gene	These class is used to represent the concept of gene.
location in terms of [...] promoters [...]	TranscriptableElement, Promoter	These class is used to represent the concept of promoters.
location in terms of [...] introns [...]	TranscriptableElement, RegulatoryElement, Intron	These class is used to represent the concept of intron.
location in terms of [...] exons [...]	TranscriptableElement, Exon	These class is used to represent the concept of Exon.
location in terms of [...] mRNAs, coding regions, and untranslated regions [...]	Transcript, PrimaryTranscript, MatureTranscript, mRNA, 5'UTR, CDS, 3'UTR	These classes are required to model the Transcripts and is compositional parts and how they are connected to the genes that codes them.
Another relevant field of study for citrus is evolutionary genomics.	OrthologousGroup	The IVIA performs evolutionary genomics by studying orthologous groups of genes among the different citrus varieties.

Table 4.5 continues on the next page

Table 4.5 continued from previous page

Sentence	Class(es)	Explanation
	Variation View	
[...] genomics regions [...]	VariationPosition	This class is required to locate variations in specific genomics regions.
[...] (i.e., SNP and small indel variations) [...]	Variation, Precise, Indel	These classes model the SNP and indel variations that are considered for this use case.
[...] notorious impact [...] the predicted impact of variations	Annotation	This class represents functional annotations used to predict the impact of variations.
one group contains varieties that highly express a characteristic of interest, and the other group contains varieties that do not express it In order to work with the varieties There are dozens of sequenced citrus varieties comparing the groups of varieties the groups of varieties have to be compared i.e., the DP and GQ values differences at a genotype level	Lecture	This class represents the varieties sequenced by the IVIA. It also contains the DP, GQ, and genotype attributes.

Table 4.5: Mapping carried out during Phase 2 (Select) of the ISGE method for generating the CSGD

After this process, all of the participants of the previous step evaluated the results and identified inconsistencies in terms of knowledge gaps (see Table 4.6). These inconsistencies were solved together with domain experts who supported this process.

Original 3–tuples	**New 3–tuple**	**Explanation**
– [Protein, *has*, Isoform], – [Isoform, *contains*, PrimaryStructureElement], – [PrimaryStructureElement, specialization, Region], – [Region, *brings_to*, SecondaryStructureElement], – [TertiaryStrutureElement, *aggregation*, SecondaryStructureElement], – [TertiaryStructureElement, *specialization*, Domain]	[Protein, *aggregation*, Domain]	The knowledge associated with the internal structure of proteins in citrus is limited to the available data, which only reports the domains that compose some of the existing proteins.
– [Transcript, *specialization*, PrimaryTranscript], – [Transcript, *specialization*, MatureTranscript] – [MatureTranscript, *specialization*, mRNA] – [Gene, *transcription*, PrimaryTranscript] – [PrimaryTranscript, *maturates_into*, MatureTranscript]	[Gene, *transcription*, mRNA]	This use case studies the protein-coding process dividing the process into two steps: from gene to mRNA, and from mRNA to protein; some of the initially selected concepts can be collapsed because they are not of interest to domain experts.
– [mRNA, *translation*, Isoform], – [Protein, *has*, Isoform],	[mRNA, *translation*, Protein]	In this use case, no protein isoforms are described.
– [RegulatoryElement, *specialization*, Intron], – [TranscriptableElement, *specialization*, Intron], – [PrimaryTransript, *aggregation*, Intron]	[ChromosomeElement, *specialization*, Intron]	Domain users want-ed to consider in- trons as simple re- gions that do not code for protein. Although they are aware of the biological imprecision of this assumption, it is more appropriate to model domain knowledge.
– [PrimaryTranscript, *aggregation*, Exon] – [PrimaryTranscript, *maturates_into*, MatureTranscript] – [MatureTranscript, *specializes*, mRNA]	[Exon, *aggregation*, mRNA]	There is no information regarding primary transcripts and how they become mature transcripts in citrus.

Table 4.6: Inconsistencies in terms of knowledge gaps for the CSGD

As a result, the original 19 3-tuples of the CSG were collapsed into five new 3-tuples. Finally, we identified one proposal for improving or enlarging the CSG. More details regarding such proposal can be seen in Table 4.7.

Conceptual view	**Proposal**	
	Title	**Description**
Conceptual Schema of the Genome for Citrus (CSGC)	Associate speciesCharacteristic with Phenotype.	Currently, the traits that characterize a species are captured by means of the speciesCharacteristic class, which is only associated with the Species class. However, such traits could be as seen complex manifestations that arise from the composition of many phenotypes. This link between both concepts is not established in the schema. We must analyze whether the SpeciesCharacteristic and Phenotype classes are two representations of the same underlying concept. If they are, they should be merged in an improved representation that captures the particularities of each representation. If they are not, they should be explicitly connected to enrich the representation of species and individuals.

Table 4.7: Proposal identified from generating the CSGC conceptual view.

3. Generate

The CSGC (see (see `07-csgc.pdf` file[8] in [180])) focuses on the appearances of a specific type of variation, the functional consequences, and the associations with pathways. The CSGC has 44 classes (see Table 4.8), with the variation, transcription, structural, and pathway views being the most important.

4.2.3 The Platform Architecture

The CitrusGenome platform is a three-tier architecture that allows domain users to effectively prioritise genomic variations potentially associated with traits of interest. This prioritisation is accomplished through advanced SNP and indel discovery via a two-step workflow.

In the first step, two groups of citrus varieties are created: one group expresses a trait of interest, and the other does not. In the second step, the symmetric difference of the variations identified in each of the two groups (based on several user-defined criteria) is obtained.

[8] `https://zenodo.org/record/7071090/files/07-csgc.pdf`

View	**Concepts**		**Completeness**
	CSG	**CSGC**	
Bibliography	10	0	0%
Location	10	2	20%
Pathway	33	12	36.36%
Phenotype	4	0	0%
Protein	19	2	10.53%
Structural	11	9	81.81%
Transcription	17	13	70.47%
Variation	11	6	54.54%
TOTAL	115	44	38.26%

Table 4.8: Number of CSG concepts considered to build the CSGC view, which supports the CitrusGenome platform.

The first tier is the database, implemented using the Oracle database management system. **The CSGC is used to derive the physical schema of the database.**

The second tier is the back end, which is implemented in JavaScript. This tier provides an API developed using the *GraphQL* query language and the *Knex* query builder library. The logic behind the analyses are carried out within this tier. To generate the desired results, **the CSGC allows for the identification of those attributes of interest that must be queried.**

The third tier is the front end, which is implemented in JavaScript using the Angular framework. The UI was developed using the GenomIUm method [193], whose phases are depicted in Fig. 4.3. GenomIUm enabled a systematic design process and a catalog of interconnected patterns that support the design of the UIs of the CitrusGenome platform (for more details regarding UI design, refer to [192]). **The GSGC is used to populate the selected GenomIUm patterns with the appropriate data.**

Similar to the Delfos Oracle Platform, there are benefits of following a conceptual model-based approach in terms of UI design. As stated above, this process was supported by the GenomIUm method.

We have selected the UI that shows the results of the analysis. Table 4.9 enumerates the specific design patterns instantiated using the CSGC, `08-citrus Genome_conceptualDesign.png` file[9] in [180] shows the corresponding concep-

[9] `https://zenodo.org/record/7071090/files/08-citrusGenome_conceptualDesign.png`

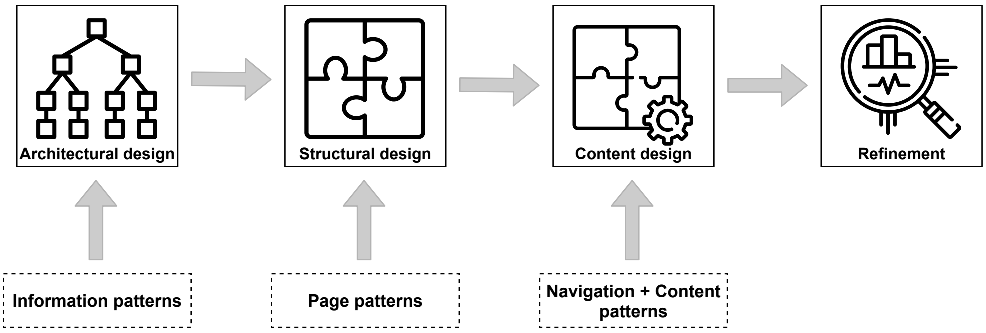

Figure 4.3: GenomIUm phases.

tual design, which is the result of applying the GenomIUm method, `09-citrusGenome_implementation.png` file[10] in [180] shows the final implementation.

ID	Pattern	Applied to	Class	Attribute(s)
1	Stepper	Guide the user through the process	—	—
2	Chart	Display the amount of identified variations	—	—
3	Ideogram	Display all chromosomes	Chromosome[]*	name
4	Chart	Show the distribution of the variations over all chromosomes	Chromosome[]*	name*
			Variation[]*	—
5	Ideogram	Display a specific chromosome	Chromosome*	—
6	Chart	Show the distribution of the variations over all chromosomes	Chromosome*	name*
			Variation[]*	—
7	Chart	Show variations distributed by variety	Variety[]*	name*
			Variation[]*	—
8	Chart	Show variations distributed by chromosome	Chromosome[]*	name*

Table 4.9 continues on the next page

[10] `https://zenodo.org/record/7071090/files/09-citrusGenome_implementation.png`

Table 4.9 continued from previous page

ID	Pattern	Applied to	Class	Attribute(s)
			Variation[]*	—
9	Chart	Show variations distributed by annotation impact	Annotation[]*	impact*
			Variation[]*	—
10	Chart	Show variations distributed by gene description	Gene[]*	description*
			Variation[]*	—
11	Chart	Show variations distributed by enzyme	Enzyme[]*	type*
			Variation[]*	—
12	Hidden Column	List, sort, and filter variations	Chromosome[]*	name*
			VariationPosition[]*	start*, end*
			Variation[]*	ref*, alt*
			ChromosomeElement[]*	name*, description*
			Entity[]*	name*, description*
			Annotation[]*	impact*, effect*, allele*
			TertiaryStructure Element[]*	name*
			Enzyme[]*	comission_number
			Event[]*	name*

Table 4.9: List of selected patterns for a specific UI in the design of CitrusGenome. The asterisk indicates mandatory data. Square brackets indicate an array of data.

4.2.4 Validation of the Platform

We validate the platform to demonstrate that these artifacts create solutions that mitigate the existing problems of genomics data management. Our goal with the validation was to determine whether the CitrusGenome platform performs efficient and user-friendly comparative analyses.

The validation comprised two phases focused on assessing the platform's ability to improve knowledge generation processes. To gather usability feedback, in the first phase, we asked domain experts to conduct a comparative analysis using CitrusGenome. In the second phase, we supervised a comparative analysis used to find variation-phenotype associations. Then, we tested the correctness and completeness of the results by comparing them to the existing body of knowledge.

First Phase

To evaluate how usable this platform is, we defined a typical genomics analysis problem to be solved by domain users using CitrusGenome. We based our evaluation on three questions (Qs):

Q1 — Will the participants be able to select the data that is involved in the genomic analysis?

Q2 — Will the participants be able to set the filters required to refine the data?

Q3 — Will the participants be able to identify the variations of interest?

The participants involved in the validation were five bioinformaticians from the IVIA. Two participants were part of the IVIA staff as senior researchers, and they had no previous experience using the platform. Two participants were Ph.D. students that work as analysts, one of whom had little previous experience using the platform while the other did not have any experience. The last participant was a Ph.D. student analyst with some previous working experience and considerable experience using the platform (see Table 4.10).

The validation consisted of 28 tasks to test the questions defined above, and it was performed virtually by using the Zoom video conference platform because of the restrictions derived from COVID-19 pandemic. On each virtual session, a participant (i.e, the **user**) carried out the proposed analysis with the support of the writer of this thesis (i.e., the **evaluator**). The evaluator guided the user

ID	Working Experience	Educational Level	Platform Experience
1	Senior	Ph.D.	None
2	Senior	Ph.D.	None
3	Medium	Ph.D. Student	Significant
4	Junior	Ph.D. Student	Little
5	Junior	Ph.D. Student	None

Table 4.10: Information of the participants that validated the UI of the Platform.

by explaining the task and asking him/her to discuss the actions or decisions made during the test.

The sessions were recorded and stored in a shared repository to enable a later review and analysis of the comments made by participants. The review and analysis process was divided into two stages. In the first stage, we met with the participants individually to review the correspondent recorded session and to get their feedback. In the second stage, we performed a focus group to discuss how easy or difficult it was for the participants to use the CitrusGenome Platform.

Participants performed the tasks related to each question, which allowed us to identify a set of usability problems. These problems constituted the basis for extending the platform in order to to overcome the difficulties the participants encountered when completing the tasks.

Set of tasks related to Q1: Will the participants be able to select the data that is involved in the genomic analysis?

All of the participants completed the tasks related to Q1 and reported a favourable opinion regarding the data selection process. They said that the drag-and-drop mechanism was intuitive and more sophisticated than other tools they previously used. However, participants 4 and 5 struggled to find two varieties because citrus varieties are identified with their common name and internal code, but these participants usually work with the variety's scientific name instead. Additionally, participant 1 noted that displaying the cultivar[11] of the varieties would enhance the process by adding relevant information that was currently missing.

[11] A cultivar is defined as a group of selected plants that share common characteristics.

Thus, the varieties' information that is shown in CitrusGenome should include both the scientific name and the cultivar. Participants 1 and 2 described potential cases where such information is necessary: i) some researchers do not know the internal code of all of the citrus varieties; and ii) the same variety can be sequenced multiple times, so the common name is not always enough to identify the variety of interest (here, the cultivar can help to select the correct one).

We defined the following usability problem from the discussions of this set of tasks:

- **Usability Problem 1**: The participant wants to know additional information (the cultivar and the scientific name) of the citrus varieties when creating the groups in the first step of the analysis.

Set of tasks related to Q2: Will the participants be able to set the filters required to refine the data?

All of the participants completed the tasks related to Q2, but those with previous experience using the platform (i.e., participants 3 and 4) completed the tasks faster than the others. All participants agreed that the way in which the filters are designed is intuitive, and they had no problem configuring them.

Participants emphasised the ease with which they set the filters, specifically with the case of the flexibility filter. The biological concept of flexibility was complex for participants 4 and 5, but the UI helped them understand this concept.

In general, all of the participants said that these filters allowed researchers to perform analyses with a high degree of modularity, flexibility, and freedom. However, the senior participants (i.e., participants 1 and 2) mentioned a missing feature they were interested in, namely, the ability to search for variations in promoter regions. This feature was of interest because promoter regions are gaining attention in recent academic work. Initially, this feature was not of interest, but this has since changed, showing that the ongoing evolution of the domain knowledge requires updating genomics tools continuously.

We defined the following usability problem from the discussions of this set of tasks:

- **Usability Problem 2**: The participant wanted to search for variations located in the promoter regions of DNA, but the option was not available.

- **Usability Problem 3**: The participant wanted to search for variations located in the promoter region of a specific gene, but the option was not available.

Set of tasks related to Q3: Will the participants be able to identify variations of interest?

The Q3 tasks are more complex and time-consuming than those of Q1 and Q2, but all participants mentioned that the UI eased the consecution of these tasks. While the participants with some experience using the platform (i.e., participants 3 and 4) mentioned that the grid provided them with a handy and flexible platform, the rest of the participants (i.e., participants 1, 2, and 5) indicated that some grid features were difficult to use.

All of the participants were able to complete the tasks except participant 5 because he did not remember which varieties corresponded to the codes displayed in the grid. As a result, participant 5 could not complete tasks 14-18 and tasks 26-28. Participant 4 also struggled to complete tasks 14-18 and tasks 26-28 because recalling the varieties that composed each group proved difficult. Thus, participants agreed that this situation should be remediated.

We defined the following usability problem from the discussions of this set of tasks:

- **Usability Problem 4**: The participants wanted to see at a glance which varieties compose each group of the analysis.
- **Usability Problem 5**: The participants had trouble identifying which of the two groups the citrus varieties belonged to. This situation complicated the identification of variety-specific values in the following grid columns: Genotype, Genotype Quality, Total Depth, Allelic Depth, and Allelic Balance.

In addition to these results, the users provided four comments that were not directly related to the research questions but are considered relevant:

Comment 1 — The participants missed a home page from which to navigate to perform the analysis. From this comment, we have defined the following usability problem:

- **Usability Problem 6**: The Platform lacks a home page.

Comment 2 — After configuring the analysis filters, participants more experienced in genomics (i.e., participants 1, 2, and 3) realised that a relevant feature was missing. It is necessary to have information of the quality of the sequencing process for citrus varieties before setting up the analysis filters. This information helps to configure certain filters (e.g., the allele balance filter) with more confidence. They said that this information should be displayed and grouped by variety and genotype. From this comment, we have defined the following usability problem:

- **Usability Problem 7**: Before starting the analysis, the participant wants to know the values of the mean value of the GQ and DP attributes of the variations of the citrus varieties based on the value of the GT attribute.

Comment 3 — Novel participants (i.e., 1, 2, 4, and 5) expressed that a tutorial would facilitate the first use of the platform. From this comment, we have defined the following usability problem:

- **Usability Problem 8**: Novel participants may struggle when using the platform for the first time.

Comment 4 — Participants aim to reuse their studies and switch between the visualisation of different analyses to compare the results. However, they can only visualise the result of an analysis when it is completed, and the information is lost if they exit the page. From this comment, we have defined the following usability problem:

- **Usability Problem 9**: The participants want to be able to visualise analyses at anytime.

Considering the validation with the user tests, we evaluated the results, identified the severity level of each usability problem, and updated the design solution to solve them in a new iteration (see Table 4.11). After the second iteration, we met with the participants of the first validation again. They were satisfied with the improvements and indicated that the platform was ready to be utilised. More details regarding the updated process can be found in [192]

ID	Usability Problem	Design Solution
1	The cultivar, common name, and scientific name of citrus varieties is required when creating the groups in the first step of the analysis.	Update the card that contains variety information in the first step of the analysis to include the needed data.
2	Participants cannot search in promoter regions of the DNA.	Allow users to search in promoter regions of the DNA.
3	Participants cannot search in the promoter region of a specific gene	Allow users to search in the promoter region of a specific gene.
4	Participants want to see which varieties compose each group of the analysis.	Implement two chart components, one per variety group, to display the required information.
5	Participant struggles differentiating between varieties in the Genotype, Genome Quality, Total Depth, Allelic Depth, and Allelic Balance columns.	Group the values of these columns into two columns, namely, Group A and Group B.
6	The platform lacks a home page.	Create a *home* page.
7	Participants wants to know the mean value of the GQ and DP attributes of the variations of the citrus varieties based on the value of the GT attribute.	Create a page displaying the required information.
8	Novel participants may struggle when using the platform for the first time.	Create a *getting started* page.
9	The participant wants to be able to visualise analyses at anytime, but it can only be seen once.	Provide users with a page to retrieve analyses at anytime using a unique id.

Table 4.11: Usability problems identified in CitrusGenome and their corresponding improvements to be implemented in the second iteration.

Second Phase

The second phase of CitrusGenome validation consisted of a SNP analysis to identify the chromosomal regions responsible for premature fruit abscission and validation of the results with the currently available body of knowledge. This analysis was performed as the final project for a student's biology bacherlo's degree at the University of Alicante (Spain) [194].

Abscission is defined as the separation of vegetative (leaves and shoots) or reproductive (flowers, flower parts, and fruits) organs from the organism. Fruit trees undergo a natural fruit abscission process to sustain an optimal number of fruits, but the expression of this process varies depending on the species. While high levels of fruit abscission reduce fruit quantity, low levels reduce fruit quality. In this context, premature fruit abscission occurs when fruits fall from the tree before their optimal harvest date. Multiple reasons can trigger premature fruit abscission, such as stress or excessive sugar accumulation, but the internal mechanisms that drive this process remain unknown.

The results of this analysis identified several regions of DNA. After comparing these regions with the current body of available knowledge, we found that one of the identified regions has a strong association with premature fruit abscission:

- Chromosome 2: There is a cluster of 108 variations near the end of this chromosome. This region contains a set of genes that code for a specific type of enzymes, called kinases, that are directly linked to abscission processes [195]–[198].

Further, the analysis reported three regions associated with processes that play an essential role in premature fruit abscission:

- Chromosome 4: There is a cluster of 549 variations that modify protein-coding genes. Seventy-four genes are affected by these variations, of which 11 are associated with processes triggered during abscission (e.g., intracellular transport, cell wall modification, synthesis of auxins, and synthesis of transcription factors) [199], [200].
- Chromosome 6: There is a cluster of 48 relevant variations affecting 19 genes. Four genes are associated with abscission processes (i.e., cell wall modification) [201], [202].
- Chromosome 9: There is a cluster of 22 variations modifying 13 genes. These genes have a relevant role in the production of auxins [203], [204].

Apart from identifying the DNA regions associated with fruit abscission, in accordance with the existing body of knowledge, the analysis also reported novel chromosomic regions that do not have supporting literatureand that can be of interest in future research:

- Chromosome 1: Between 25,500,500 and 27,000,000 bp positions, there is a cluster of 89 variations affecting fifteen genes. In total, 60 variations alter the structure of proteins

4.2.5 Conclusions

Throughout this section, we answered **RQ9**: To what extent are the contributions of this thesis useful in an agri-food genomics context? The CitrusGenome platform showed that it can efficiently perform genomic comparative studies, providing a user interface that is intuitive and easy to use. The CSGC facilitated the integration of genomics data and the definition of the filtering algorithms. In this use case, the design of the user interface was even more important due to the complexity of the analysis and the large amount of data retrieved. Again, following a conceptual model-based approach delivered high-quality user interfaces that are intuitive and easy to use.

The validation reported promising results. The first phase of the validation allowed us to validate the platform's user interface. Domain experts were very pleased with its overall quality and indicated that it was easy to use.

Going further, it showed that the platform could find variations reported in the bibliography and in novel regions. In the second phase of the validation, we compared two citrus groups. The first group comprised 17 citrus varieties without premature fruit abscission, and the second group had 12 citrus varieties with premature fruit abscission. We identified several variations located in one region with a strong association with premature fruit abscission, three regions associated with processes that play an essential role in premature fruit abscission, and one novel region without supporting literature.

Once we have finished validating the platform in a lab environment (i.e., TRL-4), CitrusGenome is ready to be tested in real-world scenarios. At the time of writing this thesis, CitrusGenome was being used by IVIA researchers to identify relevant regions of DNA that are potentially associated with specific fruit traits of interest.

4.3 Ontological Unpacking Assessment

We start this section with a brief introduction of the motivation of this experiment. Then, we describe the set-up of the experiment, report the results, and discuss them.

4.3.1 Introduction

Ontological unpacking is a procedure requiring time and effort. However, we aim to evaluate whether its benefits justify the process. Our goal is to analyse differences between representing a domain with UML and OntoUML. The purpose is to examine the pros and cons of OntoUML, with respect to domain representation.

4.3.2 Methods

We organised our empirical evaluation objectives using the Goal Question Metric template for goal definition following the guidelines for reporting software engineering experiments in [205].

Hypothesis development

According to ISO 25000 [206] usability is defined in terms of effectiveness, efficiency, and satisfaction: *"the degree to which specified users can achieve specified goals with effectiveness in use, efficiency in use and satisfaction in use in a specified context of use"*. We use this definition to specify our research questions:

- **Q1**: Is effectiveness in the conceptual modeling interpretation affected by the model notation?
- **Q2**: Is efficiency in the conceptual modeling interpretation affected by the model notation?
- **Q3**: Is satisfaction in the conceptual modeling interpretation affected by the model notation?

The research questions specified above lead to the three null hypotheses that are tested in this experiment:

- H_{01}: Effectiveness analysing a conceptual model expressed in OntoUML is the same as with UML.
- H_{02}: Efficiency analysing a conceptual model expressed in OntoUML is the same as with UML.
- H_{03}: Satisfaction analysing a conceptual model expressed in OntoUML is the same as with UML.

Factor, response variables, and metrics

In this experiment, we used the conceptual modelling notation as a factor, which presents two levels. The first level is the control treatment (i.e., the UML notation). The second level is the target treatment (i.e., the OntoUML notation). The reason for using UML as the control treatment is that it is widely known and previously used by the subjects. One response variable is defined for each null hypothesis to be tested.

The first response variable is effectiveness, defined as "the accuracy and completeness with which users achieve specified goals" [207]. We measure effectiveness through a model questionnaire composed of true/false questions that capture specific parts of the conceptual schemes. These questions are divided into three groups: related to endurants, related to perdurants, and related to the interaction between endurants and perdurants. We defined the metric per group calculated as the sum of the values associated with its answers (i.e., 1 for correct and 0 for failure).

The second response variable is efficiency, defined as "the degree to which a system or component performs its designated functions with minimum consumption of resources" [207]. We defined three metrics, each measuring the time spent for understanding the model to answer the questions on a per-group basis. The time-to-response for each group of questions is calculated as the sum of the time spent answering each question.

The third response is satisfaction, defined as "freedom from discomfort, and positive attitudes towards the use of the product" [207]. We defined three metrics: Perceived Ease of Use (PEOU), Perceived Usefulness (PU), and Intention To Use (ITU). These metrics were measured using the Method Adoption Model (MAM), which consists of a 5-point Likert scale questionnaire. The questionnaire has six PEOU questions, eight PU questions, and two ITU questions. We calculated each metric as the sum of the answers to the corresponding questions. There is one questionnaire per treatment, sharing the same template and questions.

Subjects

The experiment was conducted with a sample of twenty subjects who completed a demographic survey in order to understand their background and mitigate validity threats. Every subject was a computer engineering student in their third year. The Grade Point Average (GPA) of the students was 7.5. More than 50% of subjects (12 out of 20) had no previous working experience, and only 25% indicated that they had more than one year of working experience (mostly as junior developers).

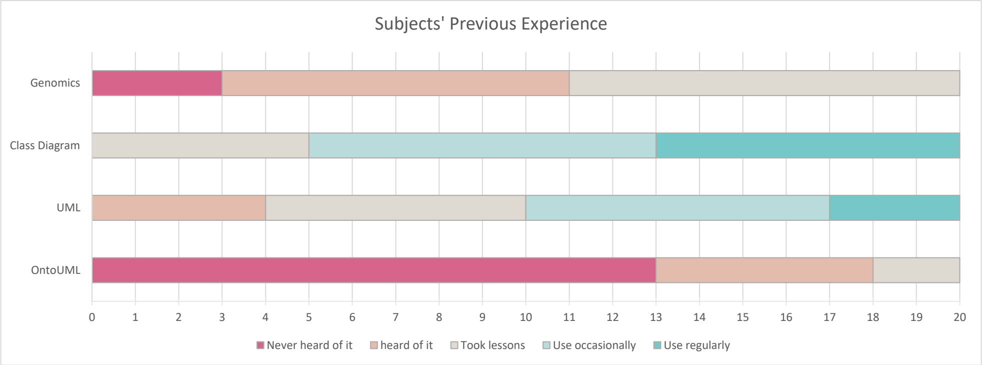

Figure 4.4: Subjects' previous experience regarding genomics, class diagrams, UML language, and OntoUML language.

Regarding the knowledge associated with the topics involved in the experiment (see Fig.4.4): all subjects knew about class diagrams and the UML language; the majority took classes in both (only four subjects did not take any class of UML); only half of the subjects took classes on genomics, whereas three of them had never heard of it; 65% of the subjects had never heard of OntoUML before the experiment, and only two of them previously studied it.

Experiment problems

The experiment aimed to asses whether OntoUML captures the particularities of the genomics domain better than UML in terms of effectiveness, efficiency, and satisfaction. To test this, we gathered relevant questions from domain experts in metabolic pathways (since the model transformation focused on the pathway view).

Problem	Group	ID	Competency Questions
P1	Entities	1	Polymers are composed of other polymers.
		2	The internal structure of any polymers is homogeneous.
		3	The internal structure of basic biological entities and polymers is the same.
	Events	4	Processes are limited in time.
		5	Pathways must be composed of other pathways.
		6	A process can be decomposed into other events.
	Interaction	7	Every biological entity must participate in at least one process.
		8	Biological entities can take part in pathways.
		9	A protein can take the roles of input, output, and regulator in the same process.
P2	Entities	10	Some polymers are composed of nucleotides.
		11	Every enzyme is a polymer.
		12	Some basic biological entities can be polymers also.
	Events	13	Every event must have a preceding event.
		14	Pathways can be composed of other pathways.
		15	Events occur in a specific time interval.
	Interaction	16	Biological entities can be created and destroyed as a result of a process.
		17	Biological entities can participate in multiple processes.
		18	A protein can take the role of input in different processes.

Table 4.12: Questions posed to subjects, clustered by Problem number and group (regarding entities, events, or their interaction).

We selected 18 questions (six associated with endurants, six associated with perdurants, and six associated with the interactions between endurants and perdurants) and distributed them between two different problems (i.e., P1 and P2). As a result, we obtained two problems with a homogeneous level of difficulty and variety (see Table 4.12 for a more detailed distribution of the questions among P1 and P2).

Group nº	First task	Second Task	Third Task	Fourth Task
1	Problem P1 (UML)	MAM (UML)	Problem P2 (OntoUML)	MAM (OntoUML)
2	Problem P2 (UML)	MAM (UML)	Problem P1 (OntoUML)	MAM (OntoUML)
3	Problem P1 (OntoUML)	PEOU-PU-ITU (OntoUML)	Problem P2 (UML)	MAM (UML)
4	Problem P2 (OntoUML)	MAM (OntoUML)	Problem P1 (UML)	MAM (UML)

Table 4.13: Organisation of the four groups for the UML-OntoUML experiment.

Experiment design

The experiment was developed as a with-in-subject design (repeated measures) in which two factors are applied to all subjects. The block variable[12] is the assigned problem because we do not aim to analyse differences between problems but if this difference affects the results. The subjects have been organised into four groups. Each group represents a possible combination of problems and treatments. Groups are balanced and subjects are randomly assigned to one group.

Experiment procedure

Once we obtained the results from the demographic survey, two teaching sessions, 45 minutes in duration, were carried out. The first session focused on the theory and practice of UML. The second session, focused on OntoUML, followed the same structure. After each class, users were asked to complete a knowledge assessment to ensure they understood the basics of UML and OntoUML, enabling the participation in the experiment. Each test was composed of eight questions that asked users about domains not associated with genomics.

Then, we distributed the participants into four groups (see Table 4.13). The first group answered P1 questions with the UML model and P2 questions with the OntoUML model. The second group answered P2 questions with the UML model and P1 questions with the OntoUML model. The third group answered P1 questions with the OntoUML model and P2 questions with the UML model. The fourth group answered P2 questions with the OntoUML model and P1 questions with the UML model.

The results were analysed based on a statistical analysis of descriptive data. We used a mixed model to identify significant differences between treatments and replications. The assumption for applying the mixed model is the normality

[12] A variable we are not interested to study but we aim to ensure that is not affecting the results.

of residuals. This can be tested when the Shapiro-Wilk test is applied to the residuals, which are automatically calculated during the application of the mixed model test [208]. The null hypothesis is rejected when the p-value is lower than 0.05, meaning the variable has significant differences.

To calculate the effect size of variables with significant differences (variables whose p-value with the mixed model is less than 0.05), we used Cohen's d [209]; it is defined as the difference between two means divided by a standard deviation of the data. According to [209], the effect size is large if Cohen's d is more than 0.8; it is moderated if it is between 0.79 and 0.5; it is small if it is between 0.49 and 0.2.

We cannot calculate power statistically (independently of the statistical tool used in the analysis) with a mixed model. However, we used G*Power to find that, for a repeated measurement statistical test, we required sample size of sixteen units for an effect size of 0.8 (large effect) to get a power of 80%. Thus, we can state that we have enough power to conduct statistical analysis with a sample of 20 units.

4.3.3 Results

Effectiveness was measured for Endurants, Perdurants, and the interaction between Endurants and Perdurants For Endurants, the results showed that OntoUML yields higher effectiveness than UML. The median, first quartile, and third quartile are significantly better for OntoUML. The results for Perdurants are very similar to the results of Endurant. However, the median, first quartile, and third quartile are almost the same for the interaction between Endurants and Perdurants.

Table 4.14 shows the statistical analysis for the metrics of effectiveness. The results of Endurants and Perdurants are significant because their p-values are lower than 0.05, and the effect size is large. Also, no significant differences have been found in the Method*Problem for these metrics, meaning that the experiment problems are not affecting the results. We can reject H_{01} for Endurants and Perdurants, concluding that the OntoUML schema is more effective than its UML counterpart for capturing the particularities of genomics.

	Treatment	**Interaction**	**Mean**	**Effect Size**
Endurants	****.001**	0.112	UML: 1.6 OntoUML: 2.3	**0.98**

Table 4.14 continued from previous page

	Treatment	Interaction	Mean	Effect Size
Perdurants	****.001**	0.388	UML: 1.7 OntoUML: 2.5	**1.2**
Interaction	0.587	0.285	UML: 1.55 OntoUML: 1.7	-

Table 4.14: Data analysis results for effectiveness metrics

Efficiency was measured for Endurants, Perdurants, and alo the interaction between Endurants and Perdurants. UML yielded higher efficiency for all. The medians, first quartile, and third quartile were higher in OntoUML.

Table 4.15 shows the statistical analysis for the metrics of effectiveness. The metrics for Endurants and Perdurants have significant results (p-value lower than 0.05) with a moderate effect size. Regarding the Method*Problem interaction, no significant results were obtained. We can reject H_{02} for Endurants and Perdurants, meaning that UML is more efficient to use. OntoUML requires significantly more usage time when compared to its UML counterpart.

	Treatment	Interaction	Mean	Effect Size
Endurants	****.006**	0.165	UML: 206.95 OntoUML: 247.65	0.4
Perdurants	****.000**	0.731	UML: 191.25 OntoUML: 251.4	0.71
Interaction	****.001**	0.468	UML: 203.65 OntoUML: 256.85	-

Table 4.15: Data analysis results for efficiency metrics

Satisfactionwas analysed through three metrics, namely, PEOU, PU, and ITU. For these metrics, the median, first quartile, and third quartile show a higher satisfaction for UML.

Table 4.16 shows the statistical analysis for the metrics of satisfaction. The three metrics showed significant results, with PEOU and ITU having a large effect size and PU having a moderate effect size. We reject H_{03} for all of the metrics; thus, UML yields significantly better satisfaction than OntoUML.

	Treatment	**Interaction**	**Mean**	**Effect Size**
PEOU	****.005**	0.843	UML: 8.5 OntoUML: 6.7	**1.1**
PU	****.003**	0.923	UML: 33.95 OntoUML: 30.9	0.78
ITU	****.005**	0.843	UML: 8.5 OntoUML: 6.7	**1.1**

Table 4.16: Data analysis results for satisfaction metrics

4.3.4 Discussion

The results showed that OntoUML is more effective than UML at capturing genomics' particularities and allowing participants to answer more questions correctly. Thus, rejecting H_{01}. This is especially the case for the cluster of questions concerning Endurants and Perdurants because we found statistical relevance. We found that Endurant-related questions were answered more successfully using OntoUML, likely because of the additional information and constraints that UFO stereotypes provided (e.g., rigidity). Perdurant-related questions were answered even more successfully with the OntoUML model. Again, this is likely due to the ontological foundation that UFO provides, capturing relevant details regarding events that remained hidden in the UML schema.

Further, a number of aspects can be discussed:

- The UML schema left implicit the fact that events have temporal time frames (i.e., they start and end at a given point in time). This information is explicitly represented in the OntoUML schema because of the constraints associated with the «event» stereotype. For instance, Q4 and Q15 focused on the temporal constraints of events, and they were answered correctly with a higher percentage with OntoUML.
- The UML schema described the participation of entities in processes in a more straightforward way, whereas its OntoUML counterpart provided a richer representation. OntoUML extended how the mereology of events are represented. For instance, it describes entities' participation in processes, which remained implicit in UML. In general, questions were answered more correctly with OntoUML. However, the increased complexity of this representation led to some questions being answered more correctly

using UML. In particular, Q6, which was associated with entities' individual participation in processes, was answered correctly with a higher percentage with OntoUML. However, Q17, associated with event mereology, was answered correctly with a higher percentage with UML. This is likely due to the increased overall complexity of events, which should be further studied.

- The «phase» stereotype allows OntoUML to exploit the principle of rigidity. This clarifies that entities can be created and destroyed due to processes. Unsurprisingly, questions involving this principle were answered with increased accuracy in OntoUML by a significant margin. For instance, Q16 interrogated participants about the principle of rigidity and more users responded correctly with OntoUML.

Regarding efficiency, we found that answering the questionnaire with OntoUML required longer response times, allowing us to reject H_{02}. Although we expected that a complex domain explained through a more complete and explicit schema would translate into reduced answering times, the results showed the contrary. These results are probably due to the increased complexity that results from grounding a conceptual schema in a foundational ontology. Moreover, participants may have needed additional time to consolidate their knowledge regarding OntoUML.

Finally, OntoUML was less appreciated by participants. The results from the MAM questionnaires allowed us to reject H_{03} since OntoUML received lower scores than UML. OntoUML was perceived as more complex, and participants were reticent to learn and use a new modeling language, especially a more complex one, in a short period.

Threats to Validity

We considered four types of threats to validity for this quasi-experiment [210], [211], namely, conclusion validity, internal validity, construct validity, and external validity.

Threats to conclusion validity prevent researchers from obtaining correct conclusions regarding the relationship between the treatment and the experiment outcome. We considered the following:

- Low statistical power: Statistical power is defined as the ability of the test to reveal a correct pattern in the data. Erroneous conclusions can be drawn when there is low statistical power. Using G*Power, we mitigated this threat by estimating the minimum sample size required to achieve statistical significance.

- Reliability of measures: The extent to which researchers can trust the measurements obtained from their experiments. Although this depends on several factors, such as poor question wording or inadequate instrumentation, the basic premise is that when measuring a phenomenon twice, the result should be the same (i.e., generate objective and reproducible measures). To mitigate this threat, we asked domain experts to double-check the list of questions for proper wording.
- Random irrelevancies in the experimental environment: These are elements outside the experimental environment that can alter the results, such as noise outside the room. We mitigated this factor by ensuring that all participants were comfortable in the classroom. They were never interrupted during the experiment, and we ensured participants did not collaborate with each other.
- Random heterogeneity of subjects: Experimental groups have a certain degree of heterogeneity. If this heterogeneity is very high, it may not be possible to assess whether the experiment results are due to individual differences or the treatment. To mitigate this threat, the experiment was conducted over a homogeneous sample of students with the same curriculum and equivalent level of knowledge regarding the class diagrams, UML, and genomics. Additionally, there were two learning lessons, one for UML and one for OntoUML, to level out possible differences among participants.

Threats to internal validity affect the experimental factor with respect to causality. Since several of such threats can be mitigated by performing a multiple-group experiment, we carried out the experiment with a four-group set-up. Then, we considered the following threat:

- Interactions with selection: This threat arises due to possible different behavior in different groups. For instance, if one group learns faster than the other. We mitigated this threat by designing an experiment in which each group applied both treatments (UML and OntoUML) to two similar problems in a different order (see Table 4.13).

Threats to construct validity can lead to results that are not generalised. Here, we considered design-related threats and social threats:

- Inadequate explanation of the constructs prior to their application: This means that the constructs are not sufficiently defined prior to being translated into measures or treatments: if the theory is not clear, neither can the experiment. This threat was mitigated by providing two learning sesions about the involved treatments (i.e., UML and OntoUML) of the same duration.
- Interaction of different treatments: When subjects participate in several studies, the treatments in these studies may interact with each other. Therefore, whether the effect is due to one of the treatments or to a combination of treatments cannot be concluded. We mitigated this threat by implementing the four-group set-up explained above.
- Restricted generalisability across constructs: This threat occurs when a treatment affects multiple constructs, some positively and others negatively. For instance, a new method improves variable A while worsening variable B; if B is not measured, a biased conclusion based on A can be drawn. To mitigate this threat, we measured three dimensions: effectiveness, efficiency, and user beliefs. Conclusions were drawn considering these.
- Evaluation comprehension: Some people are afraid of being evaluated. We tried to mitigate this threat by stating, before beginning the experiment, that no marks would be derived from this activity.
- Experimenter expectancies: Experimenters can bias the result based on the outcome they expect. Since this threat can be mitigated by involving external people without expectations, the questions were prepared by external domain experts.

Threats to external validity are those that can limit the overall generalisability of the results outside a specific context. These threats include interaction of selection and treatment, interaction of setting and treatment, and interaction of history and treatment. They are reduced by making the experimental environment as realistic as possible. Although the experiment was conducted with students, they are known to be a valid simplification of reality in a lab context [212]

4.3.5 Conclusions

Throughout this section, we answered **RQ10**: Does ontology-driven conceptual modelling capture domain particularities better than traditional conceptual modelling? The experiment results show that OntoUML captures the particularities of genomics better than UML. However, participants perceived it as difficult to use, less usefull, and reported a lower intention to use in the future. As shown in [120], [121], subjects generally need more time to understand ontology-based conceptual modelling correctly. The practical adoption of ontology-based conceptual modelling is hindered by the long learning curve associated with ontologies on the part of users. Thus, more effort should be dedicated to teaching this formalism.

Both the design of a novel OntoUML schema and the process of transforming a UML schema into its OntoUML counterpart (i.e., ontological unpacking) takes longer to achieve than developing a simpler UML schema. However, using ontology-based conceptual modeling enables better and more explicit domain representation, which can increase interdisciplinary knowledge transfer. On the one hand, domain experts should be interested in providing more precise and unambiguous representations. On the other hand, users should be interested in artifacts that ease their domain understanding by providing more precise and correct information. In conclusion, the ontological unpacking process is time-consuming, but its benefits in terms of domain understanding justify the process.

4.4 Conclusions

Throughout this chapter, we have validated the artifacts generated during the Treatment Design phase (**G3**). First, two use cases were proposed to test whether the CSG and the ISGE method deliver conceptual model-based platforms that mitigate the problems associated with managing genomics data: The Delfos Oracle and CitrusGenome. Second, an empirical experiment was conducted to test whether the ontological unpacking process benefits domain representation and understanding.

The Delfos Oracle is a platform for identifying relevant variations for a given disease. Guided by the CSG, it automates the four phases of the SILE method. This conceptual model-based solution eases genomics data integration and variation prioritisation, making knowledge generation processes more efficient and trustwothy. After validating the Delfos Oracle in a lab context with encouraging results, it is currently being tested in a real-world context to help the work of domain experts. This use case allowed us to answer **RQ7**: To what extent are the contributions of this thesis useful in a human genomics context?

The existing genomics data chaos complicates generating knowledge for precision medicine. This chaos is even worse when studying other species' genomes. The reason is that the scientific community has dedicated more effort to establishing guidelines and best practices when working with human data rather than data from other species; this is the case with citrus. In citrus, the genomics data chaos is even more significant, with fewer guidelines, standards, or established knowledge.

The CitrusGenome platform was developed to overcome the many challenges related to computational efficiency, automation, data management expertise, and the inherent limitations of data types when knowledge discovery processes are carried out when using citrus data. This platform implements a workflow for performing comparative SNP analysis involving several individuals at a whole-genome level. Also, CitrusGenome is accessible to anyone given that computer skills are not needed to use it.

Because of the complexity of the CitrusGenome workflow, we developed this tool with a strong focus on usability. Using a user-centered design supported by a sound conceptual schema allowed us to develop User Interfaces that are intuitive and easy to use. This reinforces the importance of conceptual modeling for developing high-quality user interfaces. This use case allowed us to answer **RQ8**: To what extent are the contributions of this thesis useful in an agri-food genomics context?

The human and citrus cases showed that the ISGE method eases the use of the CSG in real-world use cases and reduces the time required for its instantiation.

Although we have reported on the challenges and opportunities that arise from studying the genome, using the human and citrus genomes, we seek a much broader goal (i.e., understanding the *genome* rather than just the human genome). The development of potent genome sequencing technology, informatics, automation, and artificial intelligence, allows understanding, utilising, and conserving biodiversity like never before. For the first time, we can efficiently sequence the genomes of all known species and use genomics to help discover any remaining unknown species, of which there are many.

In the last part of Treatment Validation, we conducted an empirical experiment with 20 participants, showing that the use of ontology-based conceptual modelling results in a better representation of the domain under study. However, it also poses some challenges in terms of efficiency and user satisfaction.

Ontology-based conceptual schemes describe knowledge more explicitly and precisely, resulting in more complex representations that can require a longer time to understand. In addition, there is always a reluctance to learn new formalisms, especially when there is a long learning curve. This experiment allowed us to answer **RQ9**: Does ontology-driven conceptual modelling capture domain particularities better than traditional conceptual modelling?

Part III

Results

Chapter 5

Conclusions

We started this thesis by formulating the research questions associated with our main goals. First, we studied the problems when working with genomics data (i.e., high volume, heterogeneity, lack of interconnection, and constant evolution). Then, we studied how current approaches based on conceptualisation try to mitigate them.

Then, we applied conceptual modelling techniques to generate artifacts that improve the identified limitations. The main contributions of these thesis are the following artifacts:

1. Version 3 of the Conceptual Schema of the Human Genome.
2. The Conceptual Schema of the Citrus Genome.
3. Version 1 of the Conceptual Schema of the Genome.
4. Version 2 of the Conceptual Schema of the Genome.
5. The ISGE method.
6. The ontologically unpacked version of the pathway view of the Conceptual Schema of the Genome.

Finally, we validated the last three artifacts. The reason for not validating the first three artifacts is that are the intermediate steps that allowed us to achieve a stable conceptual schema (i.e., the CSG) ready to be used in real-world use cases. We validated the CSG and the ISGE method in use cases associated with different species, namely, humans and citrus. We developed two conceptual model-based platforms that improved the problems identified when managing genomics data. Then, we validated the ontology-based conceptual schema in an empirical experiment where twenty participants answered two questionnaires (i.e., one considering the initial UML schema and the other considering its unpacked OntoUML version).

Here, we answer to the research questions formulated in the first chapter (see Section 1.4).

5.1 Answer to Research Questions

***GOAL 1* - Study in depth the main problems in genomics data management and how they are mitigated.**

In Chapter 2, we answered the set of research questions associated with the first goal of this thesis.

RQ1 - *Which problems arise when working with genomics data?*

Working with genomics data forces geneticists and domain experts to delve into a complex ecosystem of different data sources that are not interconnected. How genomics data is represented differs significantly from one data source to another. We described the four main problems that arise when working with genomics data: first, the vast amount of data available; second, the large number of existing data sources, each of which stores data with different schemes and strategies; third, the isolation of genomics data, complicating data integration processes; fourth, the constant evolution of the genomics body of knowledge along with the data representing it. These problems are known as the genomics data chaos. We answered this question in Section 2.2.

RQ2 - *What existing approaches can be used to mitigate the identified problems?*

This thesis applied conceptual modelling techniques in the genomics domain. Thus, only those approaches that attempt to mitigate data management problems using conceptualisation as a basis were considered in our study. Two approaches based on conceptualisation are used to deal with the problems caused by the genomic data chaos. The first approach is domain ontologies. This approach provides abstract conceptualisations of particular dimensions of genomics (i.e., “vertical” dimensions) in the form of hierarchical acyclic graphs. Currently, they are the most widely adopted solution, though they have limitations regarding semantic interoperability, missing relevant parts of knowledge, or scalability. The other approach is conceptual modelling. This approach describes genomics from a holistic perspective (i.e., a “horizontal” dimension). Conceptual modelling is used to generate artifacts used in communication, understanding, and semantic interoperability. We answered this question in Section 2.3.

***GOAL 2* - Generate conceptual modelling artifacts to improve genomics data management.**

In Chapter 3, we answered the questions associated with the second goal of this thesis. Since this goal was very broad, we divided it into five subgoals: i) perform the needed updates and extensions of the Conceptual Schema of the Human Genome; ii) explore the conceptualisation of the genome for non-human species; iii) generate a conceptual schema that is species-independent; iv) provide a method to facilitate the adoption of the CSG to work with genomics use cases; and v) identify another artifacts for better representing genomics.

RQ3 - *Why/What/How to expand and update the CSHG over time?*

Genomics knowledge is constantly evolving. Thus, we must keep our conceptual schema updated to represent such an evolution of scientific knowledge. The update was conceived after analysing the feedback received from domain experts after using the Conceptual Schema of the Human Genome in multiple real-world use cases. Five dimensions that required a more precise characterisation were improved: first, to make the schema more technologically agnostic; second, to be able to represent multiple genome assemblies; third, to improve how variations are described; fourth, to model the effects caused by variations; ifth, to capture the particularities of the gene expression processes in more detail. We answered this question in Section 3.1.

RQ4 - *Why/What/How to generate a conceptual schema of the genome for non-human species?*

After working in a well-known domain such a precision medicine, we wanted to delve into the study of other species’ genomes. Although precision medicine is one of the most popular genomics fields because of its implications in human health, working domains that focus on non-human species exist. Agriculture and the improvement of agri-food crops is one example of such an application. More specifically, we focused on citrus. In this field, the amount of established knowledge is fewer when compared with human genomic data. There are also fewer standards, meaning that citrus genomic data are structured in a more technologically-oriented way. The generated conceptual schema (i.e., the Conceptual Schema of the Citrus Genome) focused on variations and their effects on proteins and metabolic pathways. All this work was performed with the *Instituto Valenciano de Investigaciones Agrarias* (IVIA). We answered this question in Section 3.2.

RQ5 - *Why/What/How to generate a conceptual schema of the genome that is species-independent?*

Genomics is a vast domain with heterogeneous use cases and one that poses several challenges. From what we have observed, generated conceptual schemes tend to focus on specific scenarios. However, every use case has a part of knowledge that is common to the other use cases, and we are not taking advantage of this. For instance, the CSHG intends to improve precision medicine and genetic diagnosis, and the CSCG focuses on supporting the identification of the genetic cause of phenotype expression in the agri-food field. Still, both conceptual schemes share a relevant portion of knowledge. To illustrate, both schemes model relevant concepts such as chromosomes or genes. Conceptual schemes that focus on a specific species could be seen as a limitation in this context because studying a different species would require creating a new conceptual schema to cover such species' particularities accurately. After thoroughly comparing our two conceptual schemes (i.e., the CSHG and the CSCG), we developed a conceptual schema that covers both use cases and is species-independent. We answered this question in Sections 3.3 and label 3.3.3.

RQ6 - *Why/What/How to create a method to generate subschemes of the Conceptual Schema of the Genome?*

We created the ISGE method to improve the adoption of the CSG. This method generates conceptual views that focus on those parts of the schema that are relevant for a specific use case, making working with the CSG more efficient. The ISGE method comprises three phases: first, identify the requirements of the use case; second, select those pieces of knowledge that are relevant for solving the identified requirements; third, generate the conceptual view with a subset of classes from the whole schema. We answered this question in Section 3.4.

RQ7 - *How to conduct ontology-driven conceptual modelling in genomics?* During our efforts to generate the needed artifacts, we found that the genomics domain has particularities that could not be captured appropriately using traditional conceptual modelling techniques. Thus, we applied ontology-based conceptual modelling techniques to assess whether they capture genomics properties better. Our effort to develop ontology-driven conceptual models in genomics has been a novel approach with encouraging results. We carried out a model-to-model transformation (i.e., the "ontological unpacking") to better capture the particularities of this complex domain. This process use a view of the CSG (i.e., the pathway view) as input and transformed it into its unpacked version. We used the UFO ontology and the OntoUML modelling language to perform such a transformation, leading to a new model that captures additional semantics with the constructs defined in UFO. This allowed us to represent critical aspects genomics in a more explicit way, leading to better domain understanding. We answered this question in Section 3.5

<u>GOAL 3</u> - Confirm and validate our contributions and research results.

In Chapter 4, we answered the questions associated with the third goal of this thesis.

RQ8 - *To what extent are the contributions of this thesis useful in a human genomic context?*

For the human genomic context, we developed a Genome Information System to improve the identification of relevant and high-quality variations (i.e., the Delfos Oracle). This platform improves variation identification by considering four data dimensions: integration, prioritisation, storage, and visualisation.

This platform took advantage of the CSG to improve the dimensions mentioned above. First, using a conceptual schema allowed us to integrate several data structures into a common schema. Second, data prioritisation rules follow a well-defined set of schema attributes. Third, the physical schema of our persistence layer is based on our the CSG. Fourth, the UI design followed a pattern-oriented approach that required a conceptual schema to be designed and implemented.

The Citrus Genome platform was validated in a lab context and showed promising results. Because of the CSG, it integrates genomics data easily and reduces the amount of relevant data associated with diseases by removing low-quality data. Additionally, its user interface is intuitive and well-designed because it follows a conceptual model-based approach. Cosnidering knowledge generation, we showed that the platform allows non-expert users to generate high-quality knowledge with high effectiveness, efficiency, and user satisfaction. We answered this question in Section 4.1.

RQ9 - *To what extent are the contributions of this thesis useful in an agri-food genomic context?*

Considering the agri-food genomic context, we developed a Genome Information System for performing efficient comparative genomic studies of DNA sequences of citrus crops (i.e., CitrusGenome). This platform took advantage of the CSG to i) integrate all of the data and ii) implement the algorithms to retrieve the data of interest. Again, a conceptual model-based approach was crucial to developing high-quality user interfaces that satisfied domain experts.

The validation of the platform was twofold. First, we asked domain experts to perform a typical analysis to evaluate this platform. Discussions where held to collect domain experts' opinions regarding CitrusGenome; the feedback was positive in terms of usability and intention to use. Second, we supervised an experiment that identified both known and novel genotype-phenotype associations for fruit abscission, a relevant trait for potentially improving crops. Apart from the three well-known regions associated with processes playing essential roles in fruit abscission that the experiment identified, an additional novel region currently being studied in more depth by domain experts was obtained. We answered this question in Section 4.2.

RQ10 - *Does ontology-driven conceptual modelling capture domain particularities better than traditional conceptual modelling?*

The OntoUML schema generated after the ontological unpacking process was confronted with its UML counterpart in an empirical experiment. The results showed that OntoUML improves the representation of several aspects of genomics, including the characterisation of biological entities, the changes in biological entities over time, and the representation of chemical compounds. These results also justify the time required to do the model transformation. We answered this question in Section 3.5.

Chapter 6

Thesis Impact

In this chapter, we present the thesis impact in terms of publications, research stays, teaching experience, participation on research projects, congress organisation, patents, and peer reviewing.

Publications

A total of 24 publications were published during while conducting the research and writing this thesis (see Table 6.1). A detailed list and summary of the publications is provided in Appendix A. There has been a high degree of interaction with other computer science and genomics experts, shown by the many researchers from different countries who co-authored these publications (i.e., 20 co-authors). We would like to highlight that the author of this thesis collaborated with:

Prof. Manuel Talon: Supervisor of the genomics center of IVIA, with almost 30,000 citations. He is an internationally recognised reference in breeding improved citrus varieties and has studied the origin of citrus. We collaborated in all the work associated with the citrus domain, including the conception of the conceptual schema of the citrus genome, the generation of the conceptual schema of the genome for citrus conceptual view using ISGE, and the development of the CitrusGenome web platform.

Prof. Stefano Ceri: Full professor at the *Politecnico di Milano* university with more than 30,000 citations. He received two advanced ERC Grants, received the ACM-SIGMOD Innovation Award in 2013, and is an ACM fellow. Our collaboration focused on validating the ISGE method to connect conceptual models inspired by different approaches (i.e., top-down and bottom-up). In addition, Dr. Anna Bernasconi, a colleague of his, stayed for six months at the Polytechnic university of Valencia. This stay led to collaborations focused on the ontological unpacking process. Future work (as explained below in Section 7) will include additional work together.

Prof. Giancarlo Guizzardi: Full professor at the University of Twente with more than 10,000 citations. He authored more than 350 publications, developed one of the most used foundational ontologies: the Unified Foundational Ontology, and is currently leading the Conceptual and Cognitive Modelling Research Group (CORE).

Category	Reference(s)	TOTAL
	JOURNALS	**7**
Q1	J2	1
Q2	J1, J3, J4, J5, J6	5
Q3	J7	1
	CONFERENCES	**14**
CORE A		10
— main	C11	*1*
— Workshop	C3, C8, C12, C13, C14	*5*
— Forum	C2, C5, C6, C9	*4*
CORE B		3
— Main	C1, C4	*2*
— Workshop	C10	*1*
CORE C		1
— Main	C7	*1*
	BOOK CHAPTERS	**3**
—	B1, B2, B3	3
	TOTAL	**24**

Table 6.1: Summary of Relevant Publications. The details of these publications can be seen in Appendix A

Finally, it is worth mentioning the following scientific results, which are under consideration for being published:

Code	Title	Year	Venue/Journal
		Journals	
X1	Usability Evaluation of a Method to Analyze Data Intensive Domains	—	Multimedia Tools and Applications (Q2)
X2	An Ontological Analysis and Assessment of Human Genome Conceptual Models	—	Journal of Biomedical Semantics (Q4)
X3	Are Automation Tools Based on the ACMG/AMP Recommendations as Homogeneous as They Should Be?	—	Briefings in Bioinformatics (Q1)

Table 6.2: Publications boing considered for publications.

Research Stay

The author of this thesis participated in a three-month research stay at the University of Twente under the supervision of Prof. Giancarlo Guizzardi. The stay allowed for intensive collaboration leading to the elaboration of the ISGE method, conception of ontological unpacking, and its first application into a relevant part of the CSG. These results materialised in multiple research outcomes: two congress publications (C9 and C14 items in Appendix A) and one journal manuscript that is being considered for publication (X2 item in Table 6.2).

Teaching Experience

The author of this thesis gained extensive teaching experience in computer science and biomedical engineering.

Computer Science

- **Computer Systems Engineering [2019, 2020, and 2021]**: Master's Degree in Software Engineering, Formal Methods, and Information Systems.

- **Information systems applied to bioinformatics: genomic data management [2022]**: Bachelor's Degree in Computer Science.
- **Information Systems Applied to Bioinformatics: Management of Genomic Data [2020, 2021, and 2022]**: Master's Degree in Software Engineering, Formal Methods, and Information Systems.
- **Model-Driven Development summer school [2022]**: 5-day summer course for French bachelor students.

Biomedical Engineering

- **Bioinformatics [2019, 2020, and 2021]**: Bachelor's Degree in Biomedical Engineering.
- **Contributions of the Biomedical Engineer [2019 and 2020]**: Bachelor's Degree in Biomedical Engineering.
- **Genomic Data Science: Towards Precision Medicine [2020]**: Course from the Permanent Training Centre in the Polythecnic University of Valencia.
- **Genomic Information Systems Design and Management [2021, 2022]**: Bachelor's Degree in Biomedical Engineering.

The author of this thesis also co-directed the following academic works:

Bachelor's Theses

- **"Analysis of Fruit Drop by Means of Identification of SNPs in Databases"**. University of Alicante, Spain, 2021. [194]
- **"Design and Development of a Web Platform for the study of Familial Cardiomyopathies"**. Polytechnic University of Valencia, Spain, 2022. [213]

Master's Thesis

- **"Identification of genomic variants using the SILE method: Extension of the research module and Focus on the Brugada Syndrome"**. Università degli Studi di Milano Bicocca, Milano, Spain, 2021. [214]

- **Design of a Conceptual Model to characterize human proteome and pathways: applications to CoVid-19 metabolism.**. Polytechnic University of Valencia, Spain, 2022. [215]

Research Projects

The author of this thesis participated in the following research projects:

- **DataMe - A Model Driven Software Production Method for Big Data Application Development** (*6 months*). Funded by the Spanish Ministry of Science and Innovation (national project). Ref: *TIN2016-80811-P.*
- **Gispro - Genomic Information Systems Production** (*32 months*). Funded by the Generalitat Valenciana (regional project). Ref: *PROMETEO/2018/176.*
- **OGMIOS - An Intelligent System for Clinical Decision Support in Precision Medicine** (*12 months*). Funded by the Generalitat Valenciana (regional project). Ref: *INNEST/2021/57.*

Congress Organisation

The author of this thesis assited with the organisation of the following congresses:

- **The 36th International Conference on Conceptual Modeling**: celebrated between November 6-9, 2017 in Valencia, Spain. *Local Organizing Committee.*
- **The 11th ACM SIGCHI Symposium on Engineering Interactive Computing Systems**: celebrated between June 18-21, 2019 in Valencia, Spain. *Local Organizing Committee.*

Patents and Software

The author of this thesis participated in the creation of following software patents:

- G-MAC: Conceptual Model-based Information System for the Effective and Efficient Management of Retina-Macula Pathology Data (reference: **S-066-2020**)
- Delfos Platform: Information System for the Management of Relevant Genomic Variations (reference: **S-075-2021**)

Peer Reviewing

The author of this thesis has peer-reviewed the following manuscripts:

- Macadamia germplasm and genomic database (MacadamiaGGD): A comprehensive platform for germplasm innovation and functional genomics in Macadamia [216].
    - *Frontiers in Plant Science* (**6,627 JIF - Q1**).

Chapter 7

Future Work

In this last chapter, we report on future work based on domain conceptualisation and the Genome Information Systems implemented to validate the contributions of this thesis. The chapter ends with some final thoughts about future lines of work that are associated with the outcomes of this thesis.

Genomics Domain Conceptual Modelling

Genomics is an ever-changing domain where new knowledge is constantly generated, and the CSG must evolve accordingly. We divided future work toward modelling genomics into three terms, based on the temporal dimension: the short term, the middle term, and the long term.

The Short Term

At this moment, we are currently preparing the next version of the CSG. Based on interactions with domain experts and clinicians, we will consider the representation of the following aspects:

- **Gene expression**: This process controls when and where RNA and proteins are generated. The expression of a gene changes dramatically depending on the surrounding environment and can regulate the expression of other genes. There are abundant heritable variations affecting gene expression [217], and their effects on diseases are a matter of study [218]. For instance, in breast cancer [219]. Representing gene expression will improve the analysis of genomics data and, potentially, deliver better, more personalised treatments.

- **Cancer**: Despite significant efforts, cancer stands as one of the major causes of morbidity and mortality. This disease appears when body cells grow uncontrollably and spread to other parts of the body. Although we already consider phenotypes and diseases in the CSG, cancer is incredibly complex, having several particularities. These particularities (e.g., the propagation type) must be represented explicitly in the model when working with cancer-associated data.

- **Somatic variations**: This is associated with the above point. Currently, the CSG considers germline variations (i.e., those variations present in every cell of the body and are inherited from parents). However, the analysis of somatic variations is critical when studying cancer. Moreover, the same variation can be germline in one individual and somatic in another; in some cases, this variation is reported with different interpretations depending on its origin (i.e., germline or somatic) [220].

- **Genome-wide association studies (GWAS)**: GWAS studies are fundamental for establishing genotype-phenotype associations [221]. These studies consist of sequencing a large number of people with a particular disease to identify common DNA variations [221]. If the identified variations are not common in a control group of healthy people, they can be considered for establishing an association with the disease of interest. These studies help develop better methods to prevent, detect, and treat diseases. However, this approach comes with challenges; most of the identified variations are located in non-coding regions (likely associated with RNA products that alter gene expression) [222]. The holistic perspective of the genome provided by conceptual modelling techniques could help extract knowledge more effectively.

- **Individuals and samples**: The CSG represents the genome from a holistic perspective, including those variations that can arise in the DNA sequence. However, it does not consider how these variations appear in individuals. Also, the particularities of each individual must be considered because the same variation can bring a different outcome from one individual to another. The samples from which DNA sequencing of individuals is carried out need to be also considered. It is relevant to identify sample aspects such as the type of sample (e.g., tissue or fluid), the sampling tool, the storage methods, or sample measures (e.g., cellularity or volume).

- Haplotypes: There are clusters of DNA variations that tend to be inherited together because they very are close to each other and are not located in recombination sites. Some phenotypes are expressed when a haplotype occurs, opening many exciting approaches to disease prevention and treatment. Also, the study of haplotypes among populations allows for the detailed study of evolutionary processes [223] and population ancestry [224].

- **Structural variations**: The CSG considers "small" DNA variations because they have been the main focus of study due to technological and economic limitations. However, larger variations that span multiple chromosomes must also be considered. These variations include rearrangements of portions of the DNA sequence from one location to another (i.e., translocations). Such variations have been associated with several diseases, such as cancer [225] or neuropathies [226], and they influence gene expression to a great extent [227].

The Middle Term

For the middle term, we refer to the study of those proposals obtained after applying the ISGE method for generating the CSGC conceptual view. In our model, we use traits and phenotypes to represent the particularities of species and their corresponding specimens. However, we must study whether these two concepts are interconnected or, even more, are two representations of the same underlying concept and could be merged. This aspect will be considered after consolidating the next version of the CSG.

The Long Term

Finally, we envision broader areas of study that will require additional efforts: epigenomics and the characterisation of prokaryotic and RNA-based organisms.

The epigenome can be seen as an information layer that is on top of the genome. It consists of chemical signals regulating gene expression [228], cell development [229], or tissue differentiation [230], among many other processes. Unlike the genome, the epigenome is dynamically altered by the environment. Also, these chemical signals can be inherited, although the extent to which this happens in humans is unclear [231]. The high relevancy of epigenomics in some areas of health and precision medicine (e.g., genetic variations associated with specific traits show epigenomic enrichment in tissues relevant to that trait, thus providing an excellent resource for understanding the molecular basis of human disease [232]) justifies its future addition in the CSG.

Working with prokaryote and RNA-based organisms will allow us to open an entirely new field of study for the CSG: the study of infectious processes and host-pathogen interactions (e.g., virus infections). The COVID pandemic has shown us the need to consider such studies. Some approaches have already demonstrated the utility of conceptual model-based approaches for improving knowledge generation [233], [234].

In this research, we applied ISGE to generate two conceptual views, one for the human case and another for the citrus case. Future work will be oriented to validate this method further. To this aim, we will search for additional domains in which the ISGE method can be used. A new dimension to this validation is being carried out [235]. In this dimension, we are using ISGE to connect two different modeling approaches. The first is concept-oriented (i.e., a top-down approach) and is represented through the CSG. The second is data-oriented (i.e., a bottom-up approach) and is represented through the Genomic Conceptual Model (GCM) [236] produced by the GeCo project [237]. The proposed connection of these approaches, by means of ISGE, reported a number of benefits. First, extensions of the conceptual schema using the input provided by real data sets are enabled for the top-down approach. Second, data records can be semantically described by high-level concepts in the bottom-up approach.

The last contribution of this thesis focused on analysing the use of ontology-based conceptual schemes for domain representation and the study of its potential benefits compared to the traditional approach. We conducted an ontological unpacking of a portion of the CSG. Future work aims to extend this transformation to the rest of the schema and study its implications in terms of effectiveness, efficiency, and user satisfaction. Further, we are starting a collaboration with researchers from the *Politecnico di Milano* University to apply ontological unpacking to their schemes, giving our work a relevant, international dimension.

Genome Information Systems

Regarding the two Genome Information Systems generated during the development of this research, there are several aspects that will be improved.

For the Delfos Oracle,we have started an industrial project to transfer our platform to the industry[1]. After finishing this project, the Delfos Oracle platform will be in the TRL–7 stage[2]. More specifically, we aim to improve all of its four modules in the following areas:

- **Hermes**: This module will benefit from including additional general-purpose and disease-specific databases. According to [238], there are 1,645 genomics databases publicly available that could potentially be included.
- **Ulises**: There are several aspects of the rule-based algorithm applied in Ulises that will be improved. We will include additional rules for considering functional and population studies when classifying variations. We will also implement support for text-based search to identify potentially relevant bibliography.

  Another aspect of Ulises that will be extended in the future is the type of variations it analyses. Currently, only SNP and INDEL variations are studied. However, we will add support to analyse additional types of variations in the future. For instance, somatic[3] and structural[4] variations.

  The final dimension of the algorithm we plan to improve is the management of non-classified variations. In the future, the standard guidelines such as the ACMG/AMP guidelines [128] will be implemented to provide our own classifications.
- **Delfos**: The information stored in the Delfos module is “static”. A manual update needs to be carried out to consider the changes over time of the information. The most important future work for this module is to implement the automatic update of the data over time.
- **Sibila**: This module will be updated to display all the new information that will be considered by the three other modules. Also, we aim to

[1] PDC2021-121243-I00 — Delfos Platform: An Information System for the Management of Genomic Variations.

[2] System prototype demonstration in operational environment, according to ISO 16290:2013.

[3] A variation occurs after conception due to environmental factors or replication errors.

[4] Variations larger than 1,000 bases. These variations can affect more than one chromosome.

improve filtering options of variations and implement a 3D visualisation of protein structure, how metabolic pathways are altered by variations, and the consequences of structural variations in chromosomes.

Regarding CitrusGenome, we are working with the IVIA to plan future work. First, we are supervising a master's thesis for implementing a concurrent data-loading system for efficiently include new citrus varieties in the system. Second, we will use the results of this master thesis to load a new set of 70 DNA sequences from citrus varieties that have been sequencing in the lab of the IVIA. These varieties will allow for extending the scope of the analyses performed with the tool. Third, we will start working with data associated with the genome of rice varieties, leading to the future development of the RiceGenome web platform.

7.1 Final considerations

This work introduced a species-independent perspective of modelling the genome. What are the consequences of such a perspective? How could this perspective determine future work? In addition to the human genome, we also modelled the genome of citrus. However, citrus is just one of the areas where high-quality DNA reference sequences have been generated. At the time of writing this thesis, several projects generated high-quality reference sequences of several species. For instance:

1. The Genome 10k project [239]: This is a consortium composed of more than 50 institutions dedicated to sample collection, genome sequencing, assembly, annotation, alignments, and analyses. As examples, the Vertebrate Genomes Project[5], aims to generate near error-free reference genome assemblies of all 66,000 extant vertebrate species, while the Earth Biogenome Project[6], aims to sequence, catalog and characterise the genomes of all of Earth's eukaryotic biodiversity.
2. The Global Invertebrate Genomics Alliance 2020 (GIGA) [240]: This is a collaborative network of researchers dedicated to building standardised best practices for sequencing invertebrates.

[5] `https://genome10k.soe.ucsc.edu/vertebrate-genomes-project/`

[6] `https://genome10k.soe.ucsc.edu/earth-bio-genome/`

3. The Darwin Tree of Life [241]: Inspired by the Earth Biogenome Project, this project is dedicated to sequencing every eukaryotic organism in Great Britain and Ireland.

Increasing our understanding of biodiversity and responsibly preserving its resources are among the most critical scientific and social challenges humans will face in the coming years [242]. Achieving such a holistic perspective will have profound benefits for us and every living being. These benefits can be grouped in three perspectives, namely, **knowledge**, **conservation**, and **health**.

Knowledge Because of the sequencing and study of other species, we can tackle fundamental questions regarding comparative biology, evolution, and genetics. For instance, in investigating chromosomal evolution among mammals [243] or detecting clade-specific conserved regions [244]. Moreover, there are dozens of topics that can be studied, such as comparative genomics of specialised traits in each vertebrate lineage, comparative genomics of convergent traits (e.g., vocal learning, flight, loss of limbs, and aquatic/terrestrial adaptations), reconstruction of common ancestors, the genetics of why some lineages are more disease resistant than others, genetic signatures of domestication across vertebrates, brain cell evolution, consequences of the evolutionary battle between transposons[7] and host factors, and many more[8].

Conservation Human activity alters the primary conditions that sustain life on Earth so significantly that biodiversity is diminished at unprecedented rates. It is our duty to conserve, protect, and even restore biodiversity; this can be achieved by increasing our knowledge about the species living on Earth. The study and sequencing of endangered and non-endangered species will increase our understanding of ecosystems; it will show how complex animal life evolved through changes in DNA, and we will be able to use this knowledge to become better stewards of the planet. The outcomes of the Genome 10k and GIGA projects will grant access to the study and conservation of species at an unprecedented scale [239].

Health The study of other species' genomes has already proved to be a valuable activity in terms of health. For instance, the first high-quality reference sequences of six bat species revealed the selection and loss of genes related to the immune system. These genes are relevant in studying emerging infectious diseases such as COVID-19 [245]. Could a better

[7] A transposon is a DNA sequence that can change its position in the genome. There are transposons that can move by themselves.

[8] `https://vertebrategenomesproject.org/`

> (or faster) response be given if such loss of genetic material had been monitored? Another exciting example is cancer, where recent discoveries indicate that tree roots penetrate soil using a mechanism similar to the one metastatic cells use to penetrate adjacent healthy tissues [246]. Such discoveries revealed that plants could be an excellent model to study how metastatic cells expand and prevent such events.

In this context, having a species-independent conceptual schema is a highly desired need, and it will allow for semantic interoperability at the genome level.

Bibliography

[1] L. Margulis *et al.*, *What Is Life?* en. University of California Press, Aug. 2000, Google-Books-ID: AbAwDwAAQBAJ, ISBN: 978-0-520-22021-8 (cit. on p. 3).

[2] A. Pross, *What is Life?: How Chemistry Becomes Biology*, en. Oxford University Press, 2016, Google-Books-ID: ETUTDAAAQBAJ, ISBN: 978-0-19-878479-1 (cit. on p. 3).

[3] C. E. Cleland *et al.*, "Defining 'Life'," en, *Origins of life and evolution of the biosphere*, vol. 32, no. 4, pp. 387–393, Aug. 2002, ISSN: 1573-0875. DOI: 10.1023/A:1020503324273 (cit. on p. 3).

[4] D. Noble, *The Music of Life: Biology beyond genes*, Inglés. Oxford, Feb. 2008, ISBN: 978-0-19-922836-2 (cit. on p. 3).

[5] L. J. Rothschild *et al.*, "Life in extreme environments," en, *Nature*, vol. 409, no. 6823, pp. 1092–1101, Feb. 2001, Number: 6823 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/35059215 (cit. on p. 3).

[6] P. Forterre, "Defining Life: The Virus Viewpoint," en, *Origins of Life and Evolution of Biospheres*, vol. 40, no. 2, pp. 151–160, Apr. 2010, ISSN: 1573-0875. DOI: 10.1007/s11084-010-9194-1 (cit. on p. 3).

[7] E. V. Koonin *et al.*, "Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question," en, *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 59, pp. 125–134, Oct. 2016, ISSN: 1369-8486. DOI: 10.1016/j.shpsc.2016.02.016 (cit. on p. 3).

[8] P. Bork, "Powers and Pitfalls in Sequence Analysis: The 70% Hurdle," en, *Genome Research*, vol. 10, no. 4, pp. 398–400, Jan. 2000, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.10.4.398 (cit. on p. 4).

[9] *A Brief Guide to Genomics*, en. [Online]. Available: `https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics` (cit. on p. 4).

[10] J. Pevsner, *Bioinformatics and Functional Genomics*, en. John Wiley & Sons, Aug. 2015, Google-Books-ID: OaRjCgAAQBAJ, ISBN: 978-1-118-58176-6 (cit. on p. 4).

[11] S. K. Burley, "An overview of structural genomics," en, *Nature Structural Biology*, vol. 7, no. 11, pp. 932–934, Nov. 2000, Number: 11 Publisher: Nature Publishing Group, ISSN: 1545-9985. DOI: `10.1038/80697` (cit. on p. 4).

[12] *Epigenomics Fact Sheet*, en. [Online]. Available: `https://www.genome.gov/about-genomics/fact-sheets/Epigenomics-Fact-Sheet` (cit. on p. 4).

[13] A. B. Gjuvsland *et al.*, "Bridging the genotype–phenotype gap: What does it take?" en, *The Journal of Physiology*, vol. 591, no. 8, pp. 2055–2066, 2013, ISSN: 1469-7793. DOI: `https://doi.org/10.1113/jphysiol.2012.248864` (cit. on p. 5).

[14] V. Orgogozo *et al.*, "The differential view of genotype–phenotype relationships," English, *Frontiers in Genetics*, vol. 6, 2015, Publisher: Frontiers, ISSN: 1664-8021. DOI: `10.3389/fgene.2015.00179` (cit. on p. 5).

[15] N. E. Navin, "Cancer genomics: One cell at a time," *Genome Biology*, vol. 15, no. 8, p. 452, Aug. 2014, ISSN: 1474-760X. DOI: `10.1186/s13059-014-0452-9` (cit. on p. 6).

[16] L. Chin *et al.*, "Cancer genomics: From discovery science to personalized medicine," en, *Nature Medicine*, vol. 17, no. 3, pp. 297–303, Mar. 2011, Number: 3 Publisher: Nature Publishing Group, ISSN: 1546-170X. DOI: `10.1038/nm.2323` (cit. on p. 6).

[17] S. Richards *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," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 17, no. 5, pp. 405–424, May 2015, ISSN: 1098-3600. DOI: `10.1038/gim.2015.30` (cit. on pp. 6, 27, 72).

[18] L. J. Stadler, "The Gene," *Science*, vol. 120, no. 3125, pp. 811–819, 1954, Publisher: American Association for the Advancement of Science, ISSN: 0036-8075. [Online]. Available: `https://www.jstor.org/stable/1681443` (cit. on p. 8).

[19] R. Falk, "What is a gene?" en, *Studies in History and Philosophy of Science Part A*, vol. 17, no. 2, pp. 133–173, Jun. 1986, ISSN: 0039-3681. DOI: `10.1016/0039-3681(86)90024-5` (cit. on p. 8).

[20] P. J. Beurton *et al.*, *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives.* Cambridge University Press, 2000 (cit. on p. 8).

[21] E. F. Keller, "The century beyond the gene," en, *Journal of Biosciences*, vol. 30, no. 1, pp. 3–10, Feb. 2005, ISSN: 0973-7138. DOI: `10.1007/BF02705144` (cit. on p. 8).

[22] H. Pearson, "What is a gene?" en, *Nature*, vol. 441, no. 7092, pp. 398–401, May 2006, Number: 7092 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: `10.1038/441398a` (cit. on p. 8).

[23] M. B. Gerstein *et al.*, "What is a gene, post-ENCODE? History and updated definition," en, *Genome Research*, vol. 17, no. 6, pp. 669–681, Jan. 2007, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.6339607 (cit. on p. 8).

[24] O. Pastor, "Conceptual Modeling of Life: Beyond the Homo Sapiens," en, in *Conceptual Modeling*, I. Comyn-Wattiau, K. Tanaka, I.-Y. Song, S. Yamamoto, and M. Saeki, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 18–31, ISBN: 978-3-319-46397-1. DOI: 10.1007/978-3-319-46397-1_2 (cit. on p. 8).

[25] D. B. T. Cox *et al.*, "RNA editing with CRISPR-Cas13," eng, *Science (New York, N.Y.)*, vol. 358, no. 6366, pp. 1019–1027, Nov. 2017, ISSN: 1095-9203. DOI: 10.1126/science.aaq0180 (cit. on p. 9).

[26] E. Pennisi, "The CRISPR Craze," en, *Science*, vol. 341, no. 6148, pp. 833–836, Aug. 2013, Publisher: American Association for the Advancement of Science Section: News Focus, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.341.6148.833 (cit. on p. 9).

[27] H. Ledford, "CRISPR, the disruptor," en, *Nature News*, vol. 522, no. 7554, p. 20, Jun. 2015, Section: News Feature. DOI: 10.1038/522020a (cit. on p. 9).

[28] O. Pastor, "Conceptual Modeling Meets the Human Genome," en, in *Conceptual Modeling - ER 2008*, Q. Li, S. Spaccapietra, E. Yu, and A. Olivé, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2008, pp. 1–11, ISBN: 978-3-540-87877-3. DOI: 10.1007/978-3-540-87877-3_1 (cit. on p. 10).

[29] O. Pastor *et al.*, "Conceptual Modeling of Human Genome: Integration Challenges," en, in *Conceptual Modelling and Its Theoretical Foundations: Essays Dedicated to Bernhard Thalheim on the Occasion of His 60th Birthday*, ser. Lecture Notes in Computer Science, A. Düsterhöft, M. Klettke, and K.-D. Schewe, Eds., Berlin, Heidelberg: Springer, 2012, pp. 231–250, ISBN: 978-3-642-28279-9. DOI: 10.1007/978-3-642-28279-9_17 (cit. on p. 10).

[30] J. F. Reyes Román *et al.*, "Applying Conceptual Modeling to Better Understand the Human Genome," en, in *Conceptual Modeling*, I. Comyn-Wattiau, K. Tanaka, I.-Y. Song, S. Yamamoto, and M. Saeki, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 404–412, ISBN: 978-3-319-46397-1. DOI: 10.1007/978-3-319-46397-1_31 (cit. on pp. 10, 11).

[31] A. M. Martínez *et al.*, "Facing the Challenges of Genome Information Systems: A Variation Analysis Prototype," en, in *Information Systems Evolution*, P. Soffer and E. Proper, Eds., ser. Lecture Notes in Business Information Processing, Berlin, Heidelberg: Springer, 2011, pp. 222–237, ISBN: 978-3-642-17722-4. DOI: 10.1007/978-3-642-17722-4_16 (cit. on p. 10).

[32] J. F. R. Román *et al.*, "GenesLove.Me: A Model-based Web-application for Direct-to-consumer Genetic Tests," Apr. 2017, pp. 133–143, ISBN: 978-989-758-250-9 (cit. on p. 10).

[33] V. Burriel *et al.*, “GeIS based on Conceptual Models for the risk assessment of Neuroblastoma,” in *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, ISSN: 2151-1357, May 2017, pp. 451–452. DOI: `10.1109/RCIS.2017.7956581` (cit. on p. 10).

[34] M. Navarrete-Hidalgo *et al.*, “Design and Implementation of a Geis for the Genomic Diagnosis using the SILE Methodology. Case Study: Congenital Cataract,” Apr. 2018, pp. 267–274, ISBN: 978-989-758-300-1. [Online]. Available: `https://www.scitepress.org/Link.aspx?doi=10.5220/0006705802670274` (cit. on p. 10).

[35] J. F. Reyes Román *et al.*, “GenesLove.Me 2.0: Improving the Prioritization of Genetic Variations,” en, in *Evaluation of Novel Approaches to Software Engineering*, E. Damiani, G. Spanoudakis, and L. A. Maciaszek, Eds., ser. Communications in Computer and Information Science, Cham: Springer International Publishing, 2018, pp. 314–333, ISBN: 978-3-030-22559-9. DOI: `10.1007/978-3-030-22559-9_14` (cit. on p. 10).

[36] J. F. R. Román *et al.*, “VarSearch: Annotating Variations using an e-Genomics Framework,” Apr. 2018, pp. 328–334, ISBN: 978-989-758-300-1 (cit. on p. 10).

[37] M. Van Der Kroon *et al.*, “Mutational Data Loading Routines for Human Genome Databases: The BRCA1 Case,” eng, *Journal of Computing Science and Engineering*, vol. 4, no. 4, pp. 291–312, 2010, Publisher: Korean Institute of Information Scientists and Engineers, ISSN: 1976-4677. DOI: `10.5626/JCSE.2010.4.4.291` (cit. on p. 10).

[38] A. L. Palacio *et al.*, “Towards an effective medicine of precision by using conceptual modelling of the genome: Short paper,” in *Proceedings of the International Workshop on Software Engineering in Healthcare Systems*, ser. SEHS ’18, New York, NY, USA: Association for Computing Machinery, May 2018, pp. 14–17, ISBN: 978-1-4503-5734-0. DOI: `10.1145/3194696.3194700` (cit. on p. 10).

[39] A. L. Palacio *et al.*, “A Method to Identify Relevant Genome Data: Conceptual Modeling for the Medicine of Precision,” en, in *Conceptual Modeling*, ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 597–609, ISBN: 978-3-030-00847-5. DOI: `10.1007/978-3-030-00847-5_44` (cit. on p. 10).

[40] A. León Palacio *et al.*, “Genomic Data Management in Big Data Environments: The Colorectal Cancer Case,” en, in *Advances in Conceptual Modeling*, C. Woo, J. Lu, Z. Li, T. W. Ling, G. Li, and M. L. Lee, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 319–329, ISBN: 978-3-030-01391-2. DOI: `10.1007/978-3-030-01391-2_36` (cit. on p. 10).

[41] C. Iñiguez-Jarrín *et al.*, “GenDomus: Interactive and Collaboration Mechanisms for Diagnosing Genetic Diseases,” Apr. 2017, pp. 91–102, ISBN: 978-989-758-250-9 (cit. on p. 10).

[42] C. Iñiguez-Jarrín *et al.*, “Defining Interaction Design Patterns to Extract Knowledge from Big Data,” en, in *Advanced Information Systems Engineering*, ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 490–504, ISBN: 978-3-319-91563-0. DOI: `10.1007/978-3-319-91563-0_30` (cit. on pp. 10, 100).

[43] C. Iñiguez-Jarrín *et al.*, "User Interface Design for Searching Biomedical Literature," *International Conference on Information Systems Development (ISD)*, Oct. 2018. [Online]. Available: `https://aisel.aisnet.org/isd2014/proceedings2018/eHealth/8` (cit. on p. 10).

[44] R. J. Wieringa, *Design Science Methodology for Information Systems and Software Engineering*, en. Springer, Nov. 2014, Google-Books-ID: xLKLBQAAQBAJ, ISBN: 978-3-662-43839-8 (cit. on pp. 11, 19).

[45] R. Dahm, "Friedrich Miescher and the discovery of DNA," en, *Developmental Biology*, vol. 278, no. 2, pp. 274–288, Feb. 2005, ISSN: 0012-1606. DOI: `10.1016/j.ydbio.2004.11.028` (cit. on p. 20).

[46] R. Dahm, "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research," en, *Human Genetics*, vol. 122, no. 6, pp. 565–581, Jan. 2008, Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 6 Publisher: Springer-Verlag, ISSN: 1432-1203. DOI: `10.1007/s00439-007-0433-0` (cit. on p. 20).

[47] "Felix Hoppe-Seyler (1825-1895) Physiological Chemist," *JAMA*, vol. 211, no. 3, pp. 493–494, Jan. 1970, ISSN: 0098-7484. DOI: `10.1001/jama.1970.03170030085015` (cit. on p. 20).

[48] M. E. Jones, "Albrecht Kossel, A Biographical Sketch," *The Yale Journal of Biology and Medicine*, vol. 26, no. 1, pp. 80–97, Sep. 1953, ISSN: 0044-0086 (cit. on p. 20).

[49] G. W. Beadle *et al.*, "Genetic Control of Biochemical Reactions in Neurospora," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 27, no. 11, pp. 499–506, Nov. 1941, ISSN: 0027-8424 (cit. on p. 20).

[50] E. Chargaff, "Chemical specificity of nucleic acids and mechanism of their enzymatic degradation," *Experientia*, vol. 6, no. 6, pp. 201–209, 1950 (cit. on p. 22).

[51] E. Chargaff, "Heraclitean fire," *Sketches from a life before Nature*, 1978 (cit. on p. 22).

[52] J. D. Watson *et al.*, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," en, *Nature*, vol. 171, no. 4356, pp. 737–738, Apr. 1953, Number: 4356 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: `10.1038/171737a0` (cit. on pp. 22, 51).

[53] S. Y. Tan *et al.*, "Severo Ochoa (1905–1993): The man behind RNA," *Singapore Medical Journal*, vol. 59, no. 1, pp. 3–4, Jan. 2018, ISSN: 0037-5675. DOI: `10.11622/smedj.2018003` (cit. on p. 23).

[54] A. Kornberg *et al.*, *DNA Replication*, Inglés. Sausalito, Calif., Jun. 2005, ISBN: 978-1-891389-44-3 (cit. on p. 23).

[55] F. Crick, "Central Dogma of Molecular Biology," en, *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970, Number: 5258 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: `10.1038/227561a0` (cit. on p. 23).

[56] M. Cobb, “60 years ago, Francis Crick changed the logic of biology,” *PLoS Biology*, vol. 15, no. 9, Sep. 2017, ISSN: 1544-9173. DOI: `10.1371/journal.pbio.2003243` (cit. on p. 23).

[57] J. Szeberényi, “The meselson-stahl experiment,” en, *Biochemistry and Molecular Biology Education*, vol. 40, no. 3, pp. 209–211, 2012, ISSN: 1539-3429. DOI: `https://doi.org/10.1002/bmb.20602` (cit. on p. 23).

[58] M. W. Nirenberg *et al.*, “The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides,” en, *Proceedings of the National Academy of Sciences*, vol. 47, no. 10, pp. 1588–1602, Oct. 1961, Publisher: National Academy of Sciences Section: PNAS Classic Article, ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.47.10.1588` (cit. on p. 24).

[59] F. Sanger *et al.*, “DNA sequencing with chain-terminating inhibitors,” en, *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, Publisher: National Academy of Sciences Section: Biological Sciences: Biochemistry, ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.74.12.5463` (cit. on p. 24).

[60] J. Shendure *et al.*, “DNA sequencing at 40: Past, present and future,” en, *Nature*, vol. 550, no. 7676, pp. 345–353, Oct. 2017, Number: 7676 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: `10.1038/nature24286` (cit. on pp. 25, 29, 31–33).

[61] M. P. Sawicki *et al.*, “Human Genome Project,” en, *The American Journal of Surgery*, vol. 165, no. 2, pp. 258–264, Feb. 1993, ISSN: 0002-9610. DOI: `10.1016/S0002-9610(05)80522-7` (cit. on p. 24).

[62] J. C. Venter *et al.*, “The sequence of the human genome,” eng, *Science (New York, N.Y.)*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001, ISSN: 0036-8075. DOI: `10.1126/science.1058040` (cit. on pp. 24, 25).

[63] T. P. Niedringhaus *et al.*, “Landscape of Next-Generation Sequencing Technologies,” *Analytical chemistry*, vol. 83, no. 12, pp. 4327–4341, Jun. 2011, ISSN: 0003-2700. DOI: `10.1021/ac2010857` (cit. on pp. 27, 31, 32).

[64] J. A. Shapiro, “Revisiting the Central Dogma in the 21st Century,” en, *Annals of the New York Academy of Sciences*, vol. 1178, no. 1, pp. 6–28, 2009, ISSN: 1749-6632. DOI: `10.1111/j.1749-6632.2009.04990.x` (cit. on pp. 27, 28).

[65] H. M. Temin *et al.*, “RNA-dependent DNA polymerase in virions of Rous sarcoma virus,” eng, *Nature*, vol. 226, no. 5252, pp. 1211–1213, Jun. 1970, ISSN: 0028-0836. DOI: `10.1038/2261211a0` (cit. on p. 27).

[66] D. Iwata-Reuyl, “An embarrassment of riches: The enzymology of RNA modification,” eng, *Current Opinion in Chemical Biology*, vol. 12, no. 2, pp. 126–133, Apr. 2008, ISSN: 1367-5931. DOI: `10.1016/j.cbpa.2008.01.041` (cit. on p. 27).

[67] A. E. House *et al.*, “Regulation of alternative splicing: More than just the ABCs,” eng, *The Journal of Biological Chemistry*, vol. 283, no. 3, pp. 1217–1221, Jan. 2008, ISSN: 0021-9258. DOI: `10.1074/jbc.R700031200` (cit. on p. 27).

[68] F. Denoeud *et al.*, “Prominent use of distal 5’ transcription start sites and discovery of a large number of additional exons in ENCODE regions,” en, *Genome Research*, vol. 17, no. 6, pp. 746–759, Jan. 2007, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.5660607 (cit. on p. 27).

[69] T. R. Gingeras, “Origin of phenotypes: Genes and transcripts,” en, *Genome Research*, vol. 17, no. 6, pp. 682–690, Jan. 2007, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.6525007 (cit. on p. 28).

[70] J. Jiricny, “The multifaceted mismatch-repair system,” eng, *Nature Reviews. Molecular Cell Biology*, vol. 7, no. 5, pp. 335–346, May 2006, ISSN: 1471-0072. DOI: 10.1038/nrm1907 (cit. on p. 28).

[71] M. Esteller, “Epigenetics in Cancer,” *New England Journal of Medicine*, vol. 358, no. 11, Mar. 2008, ISSN: 0028-4793. DOI: 10.1056/NEJMra072067 (cit. on p. 28).

[72] T. S. Dexheimer, “DNA Repair Pathways and Mechanisms,” en, in *DNA Repair of Cancer Stem Cells*, L. A. Mathews, S. M. Cabarcas, and E. M. Hurt, Eds., Dordrecht: Springer Netherlands, 2013, pp. 19–32, ISBN: 978-94-007-4590-2. DOI: 10.1007/978-94-007-4590-2_2 (cit. on p. 28).

[73] J. Gómez-Márquez, “What are the principles that govern life?” *Communicative & Integrative Biology*, vol. 13, no. 1, pp. 97–107, 2020, ISSN: 1942-0889. DOI: 10.1080/19420889.2020.1803591 (cit. on pp. 28, 29).

[74] E. R. Mardis, “A decade’s perspective on DNA sequencing technology,” *Nature*, vol. 470, no. 7333, pp. 198–203, 2011, ISSN: 00280836. DOI: 10.1038/nature09796 (cit. on p. 28).

[75] S. Goodwin *et al.*, “Coming of age: Ten years of next-generation sequencing technologies,” *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, 2016, ISSN: 14710064. DOI: 10.1038/nrg.2016.49 (cit. on p. 28).

[76] M. Y. Galperin, “The molecular biology database collection: 2008 update,” *Nucleic Acids Research*, vol. 36, no. SUPPL. 1, p. D2, Jan. 2008, ISSN: 03051048. DOI: 10.1093/nar/gkm1037 (cit. on p. 28).

[77] *The Cost of Sequencing a Human Genome*, en (cit. on p. 29).

[78] J. M. Heather *et al.*, “The sequence of sequencers: The history of sequencing DNA,” en, *Genomics*, vol. 107, no. 1, pp. 1–8, Jan. 2016, ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2015.11.003 (cit. on pp. 29–31).

[79] M. Kchouk *et al.*, “Generations of Sequencing Technologies: From First to Next Generation,” en, *Biology and Medicine*, vol. 09, no. 03, 2017, ISSN: 09748369. DOI: 10.4172/0974-8369.1000395 (cit. on pp. 30–32).

[80] D. G. Hert *et al.*, “Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods,” en, *ELECTROPHORESIS*, vol. 29, no. 23, pp. 4618–4626, 2008, ISSN: 1522-2683. DOI: https://doi.org/10.1002/elps.200800456 (cit. on p. 31).

[81] S. McGinn *et al.*, "DNA sequencing – spanning the generations," en, *New Biotechnology*, Special Issue: 15th European Congress on Biotechnology (ECB15), Istanbul, 23-26th September 2012, vol. 30, no. 4, pp. 366–372, May 2013, ISSN: 1871-6784. DOI: `10.1016/j.nbt.2012.11.012` (cit. on p. 31).

[82] C. S. Pareek *et al.*, "Sequencing technologies and genome sequencing," *Journal of Applied Genetics*, vol. 52, no. 4, pp. 413–435, 2011, ISSN: 1234-1983. DOI: `10.1007/s13353-011-0057-x` (cit. on p. 31).

[83] W. Zhou *et al.*, "A virtual sequencer reveals the dephasing patterns in error-correction code DNA sequencing," *National Science Review*, vol. 8, May 2021, ISSN: 2095-5138. DOI: `10.1093/nsr/nwaa227` (cit. on p. 31).

[84] E. E. Schadt *et al.*, "A window into third-generation sequencing," *Human Molecular Genetics*, vol. 19, no. R2, R227–R240, Oct. 2010, ISSN: 0964-6906. DOI: `10.1093/hmg/ddq416` (cit. on pp. 31, 32).

[85] V. Marx, "The big challenges of big data," en, *Nature*, vol. 498, no. 7453, pp. 255–260, Jun. 2013, ISSN: 1476-4687. DOI: `10.1038/498255a` (cit. on p. 32).

[86] Z. D. Stephens *et al.*, "Big Data: Astronomical or Genomical?" en, *PLOS Biology*, vol. 13, no. 7, e1002195, Jul. 2015, Publisher: Public Library of Science, ISSN: 1545-7885. DOI: `10.1371/journal.pbio.1002195` (cit. on p. 32).

[87] *Genomic Data Science Fact Sheet*, en. [Online]. Available: `https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science` (cit. on p. 32).

[88] E. R. Mardis, "The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic," en, *Cold Spring Harbor Perspectives in Medicine*, vol. 9, no. 9, a036269, Jan. 2019, Publisher: Cold Spring Harbor Laboratory Press, ISSN: , 2157-1422. DOI: `10.1101/cshperspect.a036269` (cit. on p. 33).

[89] A.-T. Maia *et al.*, "Big data in cancer genomics," en, *Current Opinion in Systems Biology*, Big data acquisition and analysis • Pharmacology and drug discovery, vol. 4, pp. 78–84, Aug. 2017, ISSN: 2452-3100. DOI: `10.1016/j.coisb.2017.07.007` (cit. on p. 33).

[90] T. Nawy, "Single-cell sequencing," *Nature Methods*, vol. 11, no. 1, pp. 18–18, Jan. 2014, Number: 1 Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: `10.1038/nmeth.2771` (cit. on p. 34).

[91] R. Sender *et al.*, "Revised estimates for the number of human and bacteria cells in the body," *PLOS Biology*, vol. 14, no. 8, e1002533, Aug. 19, 2016, Publisher: Public Library of Science, ISSN: 1545-7885. DOI: `10.1371/journal.pbio.1002533` (cit. on p. 34).

[92] D. J. Rigden *et al.*, "The 27th annual Nucleic Acids Research database issue and molecular biology database collection," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1–D8, Jan. 2020, ISSN: 0305-1048. DOI: `10.1093/nar/gkz1161` (cit. on p. 34).

[93] J. Davis-Turak *et al.*, "Genomics pipelines and data integration: Challenges and opportunities in the research setting," *Expert Review of Molecular Diagnostics*, vol. 17,

no. 3, pp. 225–237, Mar. 2017, ISSN: 1473-7159. DOI: `10.1080/14737159.2017.1282822` (cit. on p. 35).

[94] P. Suravajhala *et al.*, "Multi-omic data integration and analysis using systems genomics approaches: Methods and applications in animal production, health and welfare," *Genetics Selection Evolution*, vol. 48, no. 1, p. 38, Apr. 2016, ISSN: 1297-9686. DOI: `10.1186/s12711-016-0217-x` (cit. on p. 35).

[95] A. Bernasconi *et al.*, "The road towards data integration in human genomics: Players, steps and interactions," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 30–44, Jan. 2021, ISSN: 1477-4054. DOI: `10.1093/bib/bbaa080` (cit. on p. 35).

[96] S. E. Plon *et al.*, "The Ancestral Pace of Variant Reclassification," *JNCI Journal of the National Cancer Institute*, vol. 110, no. 10, pp. 1133–1134, May 2018, ISSN: 0027-8874. DOI: `10.1093/jnci/djy075` (cit. on p. 35).

[97] O. Campuzano *et al.*, "Reanalysis and reclassification of rare genetic variants associated with inherited arrhythmogenic syndromes," eng, *EBioMedicine*, vol. 54, p. 102 732, Apr. 2020, ISSN: 2352-3964. DOI: `10.1016/j.ebiom.2020.102732` (cit. on p. 35).

[98] D. T. Miller *et al.*, "ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG)," en, *Genetics in Medicine*, pp. 1–10, May 2021, Publisher: Nature Publishing Group, ISSN: 1530-0366. DOI: `10.1038/s41436-021-01172-3` (cit. on p. 35).

[99] B. Smith *et al.*, *Basic Formal Ontology for Bioinformatics.* IFOMIS Reports, 2005 (cit. on pp. 35, 39).

[100] G. Guizzardi, "Ontological foundations for structural conceptual models," English, Ph.D. Thesis, University of Twente, Oct. 2005 (cit. on pp. 35, 39).

[101] B. Smith *et al.*, *The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration*, Nov. 2007. DOI: `10.1038/nbt1346` (cit. on p. 36).

[102] S. Yon Rhee *et al.*, *Use and misuse of the gene ontology annotations*, Jul. 2008. DOI: `10.1038/nrg2363` (cit. on p. 36).

[103] S. Wu *et al.*, "Exploring the Development and Maintenance Practices in the Gene Ontology," en, *Advances in Classification Research Online*, vol. 24, no. 1, pp. 38–42, 2013, Number: 1, ISSN: 2324-9773. DOI: `10.7152/acro.v24i1.14675` (cit. on p. 37).

[104] L. Delcambre *et al.*, *A Reference Framework for Conceptual Modeling: Focusing on Conceptual Modeling Research.* Nov. 2018. DOI: `10.13140/RG.2.2.33041.07521` (cit. on p. 37).

[105] "Proceedings of the Eleventh International Conference on Data Engineering," in *Proceedings of the Eleventh International Conference on Data Engineering*, Mar. 1995. DOI: `10.1109/ICDE.1995.380416` (cit. on p. 37).

[106] T. Okayama *et al.*, "Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library.," *Bioinformatics*, vol. 14,

no. 6, pp. 472–478, Jan. 1998, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.6.472 (cit. on p. 38).

[107] C. Medigue *et al.*, "Imagene: An integrated computer environment for sequence annotation and analysis.," *Bioinformatics*, vol. 15, no. 1, pp. 2–15, Jan. 1999, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/15.1.2 (cit. on p. 38).

[108] N. W. Paton *et al.*, "Conceptual modelling of genomic information," *Bioinformatics*, vol. 16, no. 6, pp. 548–557, Jun. 2000, ISSN: 13674803. DOI: 10.1093/bioinformatics/16.6.548 (cit. on p. 38).

[109] S. Ram, "Modeling the semantics of 3d protein structures," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3288, pp. 696–708, 2004, ISSN: 03029743. DOI: 10.1007/978-3-540-30464-7_52 (cit. on p. 38).

[110] A. Bernasconi *et al.*, "Conceptual modeling for genomics: Building an integrated repository of open data," in *Conceptual Modeling*, H. C. Mayr, G. Guizzardi, H. Ma, and O. Pastor, Eds., Cham: Springer International Publishing, 2017, pp. 325–339, ISBN: 978-3-319-69904-2. DOI: 10.1007/978-3-319-69904-2_26 (cit. on p. 38).

[111] A. Bernasconi *et al.*, "META-BASE: A novel architecture for large-scale genomic metadata integration," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 543–557, Jan. 2022, Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics, ISSN: 1557-9964. DOI: 10.1109/TCBB.2020.2998954 (cit. on p. 38).

[112] A. Bernasconi *et al.*, "Empowering virus sequence research through conceptual modeling," in *Conceptual Modeling*, G. Dobbie, U. Frank, G. Kappel, S. W. Liddle, and H. C. Mayr, Eds., Cham: Springer International Publishing, 2020, pp. 388–402, ISBN: 978-3-030-62522-1. DOI: 10.1007/978-3-030-62522-1_29 (cit. on p. 38).

[113] H. L. Rehm *et al.*, "GA4gh: International policies and standards for data sharing across genomic research and healthcare," *Cell Genomics*, vol. 1, no. 2, p. 100 029, Nov. 10, 2021, ISSN: 2666-979X. DOI: 10.1016/j.xgen.2021.100029 (cit. on p. 38).

[114] J. O. B. Jacobsen *et al.*, "The GA4gh phenopacket schema: A computable representation of clinical data for precision medicine," *medRxiv*, no. 21266944, Nov. 30, 2021. DOI: 10.1101/2021.11.27.21266944 (cit. on p. 38).

[115] A. H. Wagner *et al.*, "The GA4gh variation representation specification: A computational framework for variation representation and federated identification," *Cell Genomics*, vol. 1, no. 2, p. 100 027, Nov. 10, 2021, ISSN: 2666-979X. DOI: 10.1016/j.xgen.2021.100027 (cit. on p. 38).

[116] A. Pease *et al.*, "The suggested upper merged ontology: A large ontology for the semantic web and its applications," in *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, vol. 28, 2002, pp. 7–10 (cit. on p. 39).

[117] S. de Cesare *et al.*, "BORO as a foundation to enterprise ontology," *Journal of Information Systems*, vol. 30, no. 2, pp. 83–112, Feb. 1, 2016, ISSN: 0888-7985. DOI: 10.2308/isys-51428 (cit. on p. 39).

[118] E. Bottazzi *et al.*, "Preliminaries to a dolce ontology of organisations," *International Journal of Business Process Integration and Management*, vol. 4, no. 4, pp. 225–238, 2009 (cit. on p. 39).

[119] H. Herre, "General formal ontology (GFO): A foundational ontology for conceptual modelling," in *Theory and Applications of Ontology: Computer Applications*, R. Poli, M. Healy, and A. Kameas, Eds., Dordrecht: Springer Netherlands, 2010, pp. 297–345, ISBN: 978-90-481-8847-5. DOI: 10.1007/978-90-481-8847-5_14. [Online]. Available: https://doi.org/10.1007/978-90-481-8847-5_14 (cit. on p. 39).

[120] M. Verdonck *et al.*, "Comparing traditional conceptual modeling with ontology-driven conceptual modeling: An empirical study," *Information Systems*, vol. 81, pp. 92–103, Mar. 1, 2019, ISSN: 0306-4379. DOI: 10.1016/j.is.2018.11.009 (cit. on pp. 39, 144).

[121] M. Verdonck *et al.*, "Comprehending 3d and 4d ontology-driven conceptual models: An empirical study," *Information Systems*, vol. 93, p. 101 568, Nov. 1, 2020, ISSN: 0306-4379. DOI: 10.1016/j.is.2020.101568 (cit. on pp. 39, 144).

[122] D. Kalibatiene *et al.*, "A systematic mapping with bibliometric analysis on information systems using ontology and fuzzy logic," *Applied Sciences*, vol. 11, no. 7, p. 3003, Jan. 2021, ISSN: 2076-3417. DOI: 10.3390/app11073003. (visited on 09/07/2022) (cit. on p. 39).

[123] C. M. Keet and Z. Khan, "Foundational ontologies: From theory to practice and back," *Journal of Knowledge Structures and Systems*, vol. 3, no. 1, pp. 67–71, 2022 (cit. on p. 39).

[124] R. Román and José Fabián, "Design and development of a genomic information system based on a holistic conceptual model of the human genome," Español, Accepted: 2018-03-22, Ph.D. Thesis, Polythecnic Unversity of Valencia, Mar. 2018. DOI: 10.4995/Thesis/10251/99565 (cit. on p. 42).

[125] *rs11571636 RefSNP Report - dbSNP - NCBI.* [Online]. Available: https://www.ncbi.nlm.nih.gov/snp/rs11571636 (cit. on p. 45).

[126] R. H. Waterston *et al.*, "On the sequencing of the human genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 6, pp. 3712–3716, Mar. 2002, ISSN: 00278424. DOI: 10.1073/pnas.042692499 (cit. on p. 46).

[127] Genome Reference Consortium, *GRCh38.p13 - Genome - Assembly - NCBI*, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/assembly/GCF%5C_000001405.39 (cit. on p. 46).

[128] S. Richards *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," *Genetics in Medicine*,

vol. 17, no. 5, pp. 405–424, May 2015, ISSN: 15300366. DOI: 10.1038/gim.2015.30 (cit. on pp. 48, 50, 165).

[129] ClinVar, "Representation of clinical significance in ClinVar and other variation resources at NCBI," pp. 3–6, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/ (cit. on p. 50).

[130] P. Cingolani *et al.*, "Variant annotations in VCF format," *January*, no. January, 2018, ISSN: 0305-1048. [Online]. Available: http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf (cit. on p. 50).

[131] S. R. Eddy, *Non-coding RNA genes and the modern RNA world*, Dec. 2001. DOI: 10.1038/35103511 (cit. on p. 52).

[132] Q. Nguyen *et al.*, "Expression Specificity of Disease-Associated lncRNAs: Toward Personalized Medicine," in *Current topics in microbiology and immunology*, vol. 394, Springer, 2016, pp. 237–258. DOI: 10.1007/82_2015_464 (cit. on p. 52).

[133] I. P. Michael *et al.*, "Intron retention: A common splicing event within the human kallikrein gene family," *Clinical Chemistry*, vol. 51, no. 3, pp. 506–515, Mar. 2005, ISSN: 00099147. DOI: 10.1373/clinchem.2004.042341 (cit. on p. 52).

[134] D. S. Latchman, "DNA Sequences, Transcription Factors and Chromatin Structure," *Eukaryotic Transcription Factors*, pp. 1–22, 2004. DOI: 10.1016/b978-012437178-1/50007-2 (cit. on p. 53).

[135] G. A. Wu *et al.*, "Genomics of the origin and evolution of Citrus," en, *Nature*, vol. 554, no. 7692, pp. 311–316, Feb. 2018, Number: 7692 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature25447 (cit. on p. 54).

[136] *The Genus Citrus*, en. Elsevier, 2020, ISBN: 978-0-12-812163-4. DOI: 10.1016/C2016-0-02375-6 (cit. on p. 55).

[137] P. Cingolani *et al.*, "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3," *Fly*, vol. 6, no. 2, pp. 80–92, 2012 (cit. on p. 57).

[138] R. Heinzelmann *et al.*, "Chromosomal assembly and analyses of genome-wide recombination rates in the forest pathogenic fungus Armillaria ostoyae," *Heredity*, vol. 124, no. 6, pp. 699–713, Jun. 2020, ISSN: 13652540. DOI: 10.1038/s41437-020-0306-z (cit. on p. 61).

[139] M. B. Gerstein *et al.*, *What is a gene, post-ENCODE? History and updated definition*, Jun. 2007. DOI: 10.1101/gr.6339607 (cit. on p. 62).

[140] P. Yu, D. Ma, *et al.*, "Nested genes in the human genome," *Genomics*, vol. 86, no. 4, pp. 414–422, Oct. 2005, ISSN: 08887543. DOI: 10.1016/j.ygeno.2005.06.008 (cit. on p. 62).

[141] R. H. Herai *et al.*, "Detection of human interchromosomal trans-splicing in sequence databanks," *Briefings in Bioinformatics*, vol. 11, no. 2, pp. 198–209, Mar. 2010, ISSN: 1467-5463. DOI: 10.1093/bib/bbp041 (cit. on p. 62).

[142] P. N. Campbell *et al.*, *Biochemistry illustrated : biochemistry and molecular biology in the post-genomic era*, 5. ed., re. Edinburgh: Elsevier Churchill Livingstone, 2005, p. 242, ISBN: 0443100349 (cit. on p. 62).

[143] A. Heger *et al.*, "Exhaustive enumeration of protein domain families," *Journal of Molecular Biology*, vol. 328, no. 3, pp. 749–767, May 2003, ISSN: 00222836. DOI: `10.1016/S0022-2836(03)00269-9` (cit. on p. 63).

[144] J. B. Miller *et al.*, "JustOrthologs: A fast, accurate and user-friendly ortholog identification algorithm," *Bioinformatics*, vol. 35, no. 4, pp. 546–552, 2019, ISSN: 14602059. DOI: `10.1093/bioinformatics/bty669` (cit. on p. 63).

[145] C. Chen *et al.*, "Mining of haplotype-based expressed sequence tag single nucleotide polymorphismsin citrus," *BMC Genomics*, vol. 14, no. 1, Nov. 2013, ISSN: 14712164. DOI: `10.1186/1471-2164-14-746` (cit. on p. 64).

[146] I. R. König *et al.*, "What is precision medicine?" en, *European Respiratory Journal*, vol. 50, no. 4, Oct. 2017, Publisher: European Respiratory Society Section: Reviews, ISSN: 0903-1936, 1399-3003. DOI: `10.1183/13993003.00391-2017` (cit. on p. 66).

[147] A. Agusti *et al.*, "Treatable traits: Toward precision medicine of chronic airway diseases," en, *European Respiratory Journal*, vol. 47, no. 2, pp. 410–419, Feb. 2016, Publisher: European Respiratory Society Section: Perspective, ISSN: 0903-1936, 1399-3003. DOI: `10.1183/13993003.01359-2015` (cit. on p. 66).

[148] K. P. Giese *et al.*, "The roles of protein kinases in learning and memory," en, *Learning & Memory*, vol. 20, no. 10, pp. 540–552, Jan. 2013, ISSN: 1072-0502, 1549-5485. DOI: `10.1101/lm.028449.112` (cit. on p. 67).

[149] J. P. Casas *et al.*, "C-reactive protein and coronary heart disease: A critical review," en, *Journal of Internal Medicine*, vol. 264, no. 4, pp. 295–314, 2008, ISSN: 1365-2796. DOI: `https://doi.org/10.1111/j.1365-2796.2008.02015.x` (cit. on p. 67).

[150] Leopold Jane A. *et al.*, "Emerging Role of Precision Medicine in Cardiovascular Disease," *Circulation Research*, vol. 122, no. 9, pp. 1302–1315, Apr. 2018, Publisher: American Heart Association. DOI: `10.1161/CIRCRESAHA.117.310782` (cit. on p. 67).

[151] R. Islamaj Doğan *et al.*, "Overview of the BioCreative VI Precision Medicine Track: Mining protein interactions and mutations for precision medicine," *Database*, vol. 2019, no. bay147, Jan. 2019, ISSN: 1758-0463. DOI: `10.1093/database/bay147` (cit. on p. 67).

[152] J. R. Savinainen *et al.*, "Biochemical and pharmacological characterization of the human lymphocyte antigen B-associated transcript 5 (BAT5/ABHD16A)," eng, *PloS One*, vol. 9, no. 10, e109869, 2014, ISSN: 1932-6203. DOI: `10.1371/journal.pone.0109869` (cit. on p. 68).

[153] J. D. Dombrauckas *et al.*, "Structural basis for tumor pyruvate kinase M2 allosteric regulation and catalysis," eng, *Biochemistry*, vol. 44, no. 27, pp. 9417–9429, Jul. 2005, ISSN: 0006-2960 (cit. on p. 68).

[154] D. Li *et al.*, “Precision Medicine through Antisense Oligonucleotide-Mediated Exon Skipping,” en, *Trends in Pharmacological Sciences*, vol. 39, no. 11, pp. 982–994, Nov. 2018, ISSN: 0165-6147. DOI: `10.1016/j.tips.2018.09.001` (cit. on p. 68).

[155] A. R. Kornblihtt, “Promoter usage and alternative splicing,” en, *Current Opinion in Cell Biology*, Nucleus and gene expression, vol. 17, no. 3, pp. 262–268, Jun. 2005, ISSN: 0955-0674. DOI: `10.1016/j.ceb.2005.04.014` (cit. on p. 68).

[156] O. Kelemen *et al.*, “Function of alternative splicing,” en, *Gene*, vol. 514, no. 1, pp. 1–30, Feb. 2013, ISSN: 0378-1119. DOI: `10.1016/j.gene.2012.07.083` (cit. on p. 68).

[157] C. Touriol *et al.*, “Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons,” en, *Biology of the Cell*, vol. 95, no. 3-4, pp. 169–178, 2003, ISSN: 1768-322X. DOI: `https://doi.org/10.1016/S0248-4900(03)00033-9` (cit. on p. 68).

[158] T. Chaijarasphong *et al.*, “Programmed Ribosomal Frameshifting Mediates Expression of the α-Carboxysome,” en, *Journal of Molecular Biology*, vol. 428, no. 1, pp. 153–164, Jan. 2016, ISSN: 0022-2836. DOI: `10.1016/j.jmb.2015.11.017` (cit. on p. 68).

[159] J. L. Reiter *et al.*, “A 1.8 kb alternative transcript from the human epidermal growth factor receptor gene encodes a truncated form of the receptor,” eng, *Nucleic Acids Research*, vol. 24, no. 20, pp. 4050–4056, Oct. 1996, ISSN: 0305-1048. DOI: `10.1093/nar/24.20.4050` (cit. on p. 68).

[160] W. M. T. Groenestege *et al.*, “Impaired basolateral sorting of pro-EGF causes isolated recessive renal hypomagnesemia,” eng, *The Journal of Clinical Investigation*, vol. 117, no. 8, pp. 2260–2267, Aug. 2007, ISSN: 0021-9738. DOI: `10.1172/JCI31680` (cit. on p. 68).

[161] T. Sanavia *et al.*, “Limitations and challenges in protein stability prediction upon genome variations: Towards future applications in precision medicine,” en, *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1968–1979, Jan. 2020, ISSN: 2001-0370. DOI: `10.1016/j.csbj.2020.07.011` (cit. on p. 68).

[162] L. R. Engelking, *Textbook of Veterinary Physiological Chemistry*, en. Elsevier, 2015, ISBN: 978-0-12-391909-0. DOI: `10.1016/C2010-0-66047-0` (cit. on p. 68).

[163] F. A. Witzmann *et al.*, “Pharmacoproteomics in drug development,” en, *The Pharmacogenomics Journal*, vol. 3, no. 2, pp. 69–76, Jan. 2003, Number: 2 Publisher: Nature Publishing Group, ISSN: 1473-1150. DOI: `10.1038/sj.tpj.6500164` (cit. on p. 68).

[164] The UniProt Consortium, “UniProt: A worldwide hub of protein knowledge,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, ISSN: 0305-1048. DOI: `10.1093/nar/gky1049` (cit. on p. 68).

[165] W. T. Godbey, “Chapter 2 - Proteins,” en, in *An Introduction to Biotechnology*, W. T. Godbey, Ed., Woodhead Publishing, Jan. 2014, pp. 9–33, ISBN: 978-1-907568-28-2. DOI: `10.1016/B978-1-907568-28-2.00002-2` (cit. on pp. 70, 71).

[166] L. P. Tripathi *et al.*, “Network-Based Analysis for Biological Discovery,” en, in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov,

K. Nakai, and C. Schönbach, Eds., Oxford: Academic Press, Jan. 2019, pp. 283–291, ISBN: 978-0-12-811432-2. DOI: 10.1016/B978-0-12-809633-8.20674-2 (cit. on p. 71).

[167] K. Nykamp *et al.*, "Sherloc: A comprehensive refinement of the ACMG–AMP variant classification criteria," en, *Genetics in Medicine*, vol. 19, no. 10, pp. 1105–1117, Oct. 2017, Number: 10 Publisher: Nature Publishing Group, ISSN: 1530-0366. DOI: 10.1038/gim.2017.37 (cit. on p. 72).

[168] A. León, A. García S., *et al.*, "Evolution of an Adaptive Information System for Precision Medicine," en, in *Intelligent Information Systems*, S. Nurcan and A. Korthaus, Eds., vol. 424, Series Title: Lecture Notes in Business Information Processing, Cham: Springer International Publishing, 2021, pp. 3–10, ISBN: 978-3-030-79108-7. DOI: 10.1007/978-3-030-79108-7_1 (cit. on p. 72).

[169] *Clinvar variant details (vcv000039614.14)*, https://www.ncbi.nlm.nih.gov/clinvar/variation/39614/, Accessed: 2020-10-21 (cit. on p. 74).

[170] S. M. Harrison *et al.*, "Using ClinVar as a Resource to Support Variant Interpretation," *Current Protocols in Human Genetics*, vol. 89, no. 1, pp. 8.16.1–8.16.23, Apr. 2016, ISSN: 1934-8266. DOI: 10.1002/0471142905.hg0816s89 (cit. on p. 75).

[171] G. John *et al.*, "Natural monomers: A mine for functional and sustainable materials – Occurrence, chemical modification and polymerization," en, *Progress in Polymer Science*, vol. 92, pp. 158–209, May 2019, ISSN: 0079-6700. DOI: 10.1016/j.progpolymsci.2019.02.008 (cit. on p. 77).

[172] N. M. Delzenne, "Oligosaccharides: State of the art," en, *Proceedings of the Nutrition Society*, vol. 62, no. 1, pp. 177–182, Feb. 2003, Publisher: Cambridge University Press, ISSN: 1475-2719, 0029-6651. DOI: 10.1079/PNS2002225 (cit. on p. 77).

[173] A. A. Zamyatnin *et al.*, "The EROP-Moscow oligopeptide database," *Nucleic Acids Research*, vol. 34, no. suppl_1, pp. D261–D266, Jan. 2006, ISSN: 0305-1048. DOI: 10.1093/nar/gkj008 (cit. on p. 77).

[174] X. Zhang *et al.*, "Oligodendroglial glycolytic stress triggers inflammasome activation and neuropathology in Alzheimer's disease," en, *Science Advances*, vol. 6, no. 49, eabb8680, Dec. 2020, Publisher: American Association for the Advancement of Science Section: Research Article, ISSN: 2375-2548. DOI: 10.1126/sciadv.abb8680 (cit. on p. 79).

[175] J. V. Pluvinage and othres, "CD22 blockade restores homeostatic microglial phagocytosis in ageing brains," en, *Nature*, vol. 568, no. 7751, pp. 187–192, Apr. 2019, Number: 7751 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-019-1088-4 (cit. on p. 79).

[176] J. A. Fifita *et al.*, "A novel amyotrophic lateral sclerosis mutation in OPTN induces ER stress and Golgi fragmentation in vitro," eng, *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration*, vol. 18, no. 1-2, pp. 126–133, Feb. 2017, ISSN: 2167-9223. DOI: 10.1080/21678421.2016.1218517 (cit. on p. 79).

[177] J. Vittitow *et al.*, “Expression of optineurin, a glaucoma-linked gene, is influenced by elevated intraocular pressure,” eng, *Biochemical and Biophysical Research Communications*, vol. 298, no. 1, pp. 67–74, Oct. 2002, ISSN: 0006-291X. DOI: `10.1016/s0006-291x(02)02395-1` (cit. on p. 79).

[178] T. Rezaie *et al.*, “Adult-onset primary open-angle glaucoma caused by mutations in optineurin,” eng, *Science (New York, N.Y.)*, vol. 295, no. 5557, pp. 1077–1079, Feb. 2002, ISSN: 1095-9203. DOI: `10.1126/science.1066901` (cit. on p. 79).

[179] T. K. Chaudhuri *et al.*, “Protein-misfolding diseases and chaperone-based therapeutic approaches,” eng, *The FEBS journal*, vol. 273, no. 7, pp. 1331–1349, Apr. 2006, ISSN: 1742-464X. DOI: `10.1111/j.1742-4658.2006.05181.x` (cit. on p. 79).

[180] A. García S., “Ph.d. thesis supporting material,” Ph.D. dissertation, Polytechnic University of Valencia, Sep. 2022. DOI: `10.5281/zenodo.7070548`. [Online]. Available: `https://doi.org/10.5281/zenodo.7070548` (cit. on pp. 85, 89, 100, 105, 108, 115, 121–123).

[181] J. F. Allen, “Maintaining knowledge about temporal intervals,” *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, Nov. 1, 1983, ISSN: 0001-0782. DOI: `10.1145/182.358434` (cit. on p. 91).

[182] G. Guizzardi *et al.*, “Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story,” *Applied Ontology*, vol. 10, no. 3, pp. 259–271, Jan. 1, 2015, ISSN: 1570-5838. DOI: `10.3233/AO-150157` (cit. on p. 91).

[183] J. S. Hamid *et al.*, “Data integration in genetics and genomics: Methods and challenges,” *Human Genomics and Proteomics : HGP*, vol. 2009, p. 869 093, Jan. 12, 2009, ISSN: 1757-4242. DOI: `10.4061/2009/869093` (cit. on p. 99).

[184] A. L. Palacio and Ó. P. López, “Smart Data for Genomic Information Systems: The SILE Method,” en-US, *Complex Systems Informatics and Modeling Quarterly*, vol. 0, no. 17, pp. 1–23, Dec. 2018, Number: 17, ISSN: 2255-9922. DOI: `10.7250/csimq.2018-17.01` (cit. on p. 99).

[185] R. Cacace *et al.*, “Molecular genetics of early-onset alzheimer’s disease revisited,” *Alzheimer’s & Dementia*, vol. 12, no. 6, pp. 733–748, Jun. 1, 2016, ISSN: 1552-5260. DOI: `10.1016/j.jalz.2016.01.012` (cit. on p. 109).

[186] M. Costa Sánchez, “Diseño y aplicación de un método basado en sile para la identificación de variaciones genéticas relevantes asociadas a la enfermedad de alzheimer temprano,” Master’s Thesis, Universitat Politècnica de València, 2020 (cit. on p. 109).

[187] K. M. Flanigan, “Duchenne and becker muscular dystrophies,” *Neurologic Clinics*, vol. 32, no. 3, pp. 671–688, Aug. 1, 2014, Publisher: Elsevier, ISSN: 0733-8619, 1557-9875. DOI: `10.1016/j.ncl.2014.05.002` (cit. on p. 109).

[188] E. M. López García, “Método de identificación de variaciones genéticas relevantes asociadas a la distrofia muscular de duchenne y de becker basado en la inteligencia artificial explicable,” Bachelor’s Thesis, Universitat Politècnica de València, Jul. 2021.

[Online]. Available: https://riunet.upv.es:443/handle/10251/171055 (cit. on p. 109).

[189] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989, ISSN: 0276-7783. DOI: 10.2307/249008 (cit. on p. 111).

[190] D. Moody *et al.*, "Evaluating the quality of information models: Empirical testing of a conceptual model quality framework," in *25th International Conference on Software Engineering, 2003. Proceedings.*, ISSN: 0270-5257, May 2003, pp. 295–305. DOI: 10.1109/ICSE.2003.1201209 (cit. on p. 111).

[191] G. A. Wu *et al.*, "Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication," *Nature Biotechnology*, vol. 32, no. 7, pp. 656–662, Jun. 2014, Publisher: Nature Publishing Group, ISSN: 15461696. DOI: 10.1038/nbt.2906 (cit. on p. 113).

[192] A. García S. *et al.*, "CitrusGenome: Applying user centered design for evaluating the usability of genomic user interfaces," in *Evaluation of Novel Approaches to Software Engineering*, R. Ali, H. Kaindl, and L. A. Maciaszek, Eds., ser. Communications in Computer and Information Science, Cham: Springer International Publishing, 2022, pp. 213–240, ISBN: 978-3-030-96648-5. DOI: 10.1007/978-3-030-96648-5_10 (cit. on pp. 115, 122, 129).

[193] C. Iñiguez-Jarrin, "GenomIUm: A pattern based method for designing user interfaces for genomic data access," Ph.D. dissertation, Universitat Politècnica de València, 2019, 193 pp. (cit. on p. 122).

[194] J. Llorens, "Analysis of fruit drop by means of identification of snps in databases," Bachelor's Thesis, University of Alicante, 2021 (cit. on pp. 130, 158).

[195] S. K. Cho *et al.*, "Regulation of floral organ abscission in Arabidopsis thaliana," en, *Proceedings of the National Academy of Sciences*, vol. 105, no. 40, pp. 15 629–15 634, Oct. 2008, Publisher: National Academy of Sciences Section: Biological Sciences, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0805539105 (cit. on p. 131).

[196] M. W. Lewis *et al.*, "The SERK1 receptor-like kinase regulates organ separation in Arabidopsis flowers," eng, *The Plant Journal: For Cell and Molecular Biology*, vol. 62, no. 5, pp. 817–828, Jun. 2010, ISSN: 1365-313X. DOI: 10.1111/j.1365-313X.2010.04194.x (cit. on p. 131).

[197] C. A. Burr *et al.*, "CAST AWAY, a Membrane-Associated Receptor-Like Kinase, Inhibits Organ Abscission in Arabidopsis," *Plant Physiology*, vol. 156, no. 4, pp. 1837–1850, Aug. 2011, ISSN: 0032-0889. DOI: 10.1104/pp.111.175224 (cit. on p. 131).

[198] J. Corbacho *et al.*, "Transcriptomic Events Involved in Melon Mature-Fruit Abscission Comprise the Sequential Induction of Cell-Wall Degrading Genes Coupled to a Stimulation of Endo and Exocytosis," en, *PLOS ONE*, vol. 8, no. 3, e58363, Mar. 2013, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0058363 (cit. on p. 131).

[199] P. Merelo *et al.*, "Cell Wall Remodeling in Abscission Zone Cells during Ethylene-Promoted Fruit Abscission in Citrus," eng, *Frontiers in Plant Science*, vol. 8, p. 126, 2017, ISSN: 1664-462X. DOI: 10.3389/fpls.2017.00126 (cit. on p. 131).

[200] L. H. Estornell *et al.*, "Elucidating mechanisms underlying organ abscission," eng, *Plant Science: An International Journal of Experimental Plant Biology*, vol. 199-200, pp. 48–60, Feb. 2013, ISSN: 1873-2259. DOI: 10.1016/j.plantsci.2012.10.008 (cit. on p. 131).

[201] S. Cai *et al.*, "Stamen abscission zone transcriptome profiling reveals new candidates for abscission control: Enhanced retention of floral organs in transgenic plants over-expressing Arabidopsis ZINC FINGER PROTEIN2," eng, *Plant Physiology*, vol. 146, no. 3, pp. 1305–1321, Mar. 2008, ISSN: 0032-0889. DOI: 10.1104/pp.107.110908 (cit. on p. 131).

[202] D. Dietrich *et al.*, "AtPTR1, a plasma membrane peptide transporter expressed during seed germination and in vascular tissue of Arabidopsis," eng, *The Plant Journal: For Cell and Molecular Biology*, vol. 40, no. 4, pp. 488–499, Nov. 2004, ISSN: 0960-7412. DOI: 10.1111/j.1365-313X.2004.02224.x (cit. on p. 131).

[203] C.-Y. Zhao *et al.*, "PI4Kγ2 Interacts with E3 Ligase MIEL1 to Regulate Auxin Metabolism and Root Development1," *Plant Physiology*, vol. 184, no. 2, pp. 933–944, Oct. 2020, ISSN: 0032-0889. DOI: 10.1104/pp.20.00799 (cit. on p. 131).

[204] J. E. Taylor *et al.*, "Signals in abscission," en, *New Phytologist*, vol. 151, no. 2, pp. 323–340, 2001, ISSN: 1469-8137. DOI: 10.1046/j.0028-646x.2001.00194.x (cit. on p. 131).

[205] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering: An Introduction.* Berlin Heidelberg: Springer, 2012 (cit. on p. 133).

[206] ISO/IEC, "Iso/iec 25000 - software engineering - software product quality requirements and evaluation (square) - guide to square," 2010 (cit. on p. 133).

[207] IEEE, *IEEE standard computer dictionary. A compilation of IEEE standard computer glossaries.* Institute of Electrical and Electronics Engineers. New York, EE.UU., 1991 (cit. on p. 134).

[208] L. S. Meyers, G. Gamst, and A. Guarino, *Applied multivariate research : design and interpretation.* Thousand Oaks, California: SAGE Publications, 2006. [Online]. Available: http://www.loc.gov/catdir/toc/ecip0510/2005009519.html (cit. on p. 138).

[209] L. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd. Edition. New York, New York: Lawrence Earlbaum Associates, 1988 (cit. on p. 138).

[210] D. T. Campbell, "Experimental and quasi-experimental designs for research on teaching," *Handbook of research on teaching*, vol. 5, pp. 171–246, 1963 (cit. on p. 141).

[211] T. D. Cook, D. T. Campbell, and A. Day, *Quasi-experimentation: Design & analysis issues for field settings.* Houghton Mifflin Boston, 1979, vol. 351 (cit. on p. 141).

[212] D. Falessi, N. Juristo, C. Wohlin, *et al.*, “Empirical software engineering experts on the use of students and professionals in experiments,” *Empirical Software Engineering*, vol. 23, no. 1, pp. 452–489, 2018 (cit. on p. 143).

[213] A. Pérez, “Design and development of a web platform for the study of familial cardiomyopathies,” Bachelor’s Thesis, Polytechnic University of Valencia, 2022 (cit. on p. 158).

[214] M. Ventola, “Identification of genomic variants using the sile method: Extension of the research module and focus on the brugada syndrome,” Master’s Thesis, University of Bicocca, 2021 (cit. on p. 158).

[215] L. Romero, “Design of a conceptual model to characterize human proteome and pathways: Applications to covid-19 metabolism.,” Master’s Thesis, Polytechnic University of Valencia, 2022 (cit. on p. 159).

[216] P. Wang *et al.*, “Macadamia germplasm and genomic database (MacadamiaGGD): A comprehensive platform for germplasm innovation and functional genomics in macadamia,” *Frontiers in Plant Science*, vol. 13, 2022, ISSN: 1664-462X (cit. on p. 160).

[217] M. V. Rockman *et al.*, “Genetics of global gene expression,” *Nature Reviews Genetics*, vol. 7, no. 11, pp. 862–872, Nov. 2006, Number: 11 Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: `10.1038/nrg1964` (cit. on p. 162).

[218] V. Emilsson *et al.*, “Genetics of gene expression and its effect on disease,” *Nature*, vol. 452, no. 7186, pp. 423–428, Mar. 2008, Number: 7186 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: `10.1038/nature06758` (cit. on p. 162).

[219] C. Sotiriou *et al.*, “Gene-expression signatures in breast cancer,” *New England Journal of Medicine*, vol. 360, no. 8, pp. 790–800, Feb. 19, 2009, ISSN: 0028-4793. DOI: `10.1056/NEJMra0801289` (cit. on p. 162).

[220] E. W. Moody *et al.*, “Comparison of somatic and germline variant interpretation in hereditary cancer genes,” *JCO Precision Oncology*, no. 3, pp. 1–8, Dec. 2019, Publisher: Wolters Kluwer. DOI: `10.1200/PO.19.00144` (cit. on p. 162).

[221] M. M. Iles, “What can genome-wide association studies tell us about the genetics of common disease?” *PLOS Genetics*, vol. 4, no. 2, e33, Feb. 29, 2008, Publisher: Public Library of Science, ISSN: 1553-7404. DOI: `10.1371/journal.pgen.0040033` (cit. on p. 162).

[222] B. Liu *et al.*, “Abundant associations with gene expression complicate GWAS follow-up,” *Nature Genetics*, vol. 51, no. 5, pp. 768–769, May 2019, Number: 5 Publisher: Nature Publishing Group, ISSN: 1546-1718. DOI: `10.1038/s41588-019-0404-0` (cit. on p. 162).

[223] M. Leitwein *et al.*, “Using haplotype information for conservation genomics,” *Trends in Ecology & Evolution*, vol. 35, no. 3, pp. 245–258, Mar. 1, 2020, ISSN: 0169-5347. DOI: `10.1016/j.tree.2019.10.012` (cit. on p. 163).

[224] E. Gilbert *et al.*, "Revealing the recent demographic history of europe via haplotype sharing in the UK biobank," *Proceedings of the National Academy of Sciences*, vol. 119, no. 25, e2119281119, Jun. 21, 2022, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.2119281119 (cit. on p. 163).

[225] K. Yi *et al.*, "Patterns and mechanisms of structural variations in human cancer," *Experimental & Molecular Medicine*, vol. 50, no. 8, pp. 1–11, Aug. 2018, Number: 8 Publisher: Nature Publishing Group, ISSN: 2092-6413. DOI: 10.1038/s12276-018-0112-3 (cit. on p. 163).

[226] A. N. Cutrupi *et al.*, "Structural variations causing inherited peripheral neuropathies: A paradigm for understanding genomic organization, chromatin interactions, and gene dysregulation," *Molecular Genetics & Genomic Medicine*, vol. 6, no. 3, pp. 422–433, 2018, ISSN: 2324-9269. DOI: 10.1002/mgg3.390 (cit. on p. 163).

[227] M. Spielmann *et al.*, "Structural variations, the regulatory landscape of the genome and their alteration in human disease," *BioEssays*, vol. 35, no. 6, pp. 533–543, 2013, ISSN: 1521-1878. DOI: 10.1002/bies.201200178 (cit. on p. 163).

[228] E. R. Gibney *et al.*, "Epigenetics and gene expression," *Heredity*, vol. 105, no. 1, pp. 4–13, Jul. 2010, Number: 1 Publisher: Nature Publishing Group, ISSN: 1365-2540. DOI: 10.1038/hdy.2010.54 (cit. on p. 163).

[229] H. a. Sasaki, "Epigenetic events in mammalian germ-cell development: Reprogramming and beyond," *Nature Reviews Genetics*, vol. 9, no. 2, pp. 129–140, Feb. 2008, Number: 2 Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: 10.1038/nrg2295 (cit. on p. 163).

[230] T. Sugimura *et al.*, "Genetic and epigenetic alterations in carcinogenesis," *Mutation Research/Reviews in Mutation Research*, vol. 462, no. 2, pp. 235–246, Apr. 1, 2000, ISSN: 1383-5742. DOI: 10.1016/S1383-5742(00)00005-3 (cit. on p. 163).

[231] E. Heard *et al.*, "Transgenerational epigenetic inheritance: Myths and mechanisms," *Cell*, vol. 157, no. 1, pp. 95–109, Mar. 27, 2014, ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.02.045 (cit. on p. 164).

[232] A. Kundaje *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, Feb. 2015, Number: 7539 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature14248 (cit. on p. 164).

[233] A. Bernasconi *et al.*, "A conceptual model for geo-online exploratory data visualization: The case of the COVID-19 pandemic," *Information*, vol. 12, no. 2, p. 69, Feb. 2021, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2078-2489. DOI: 10.3390/info12020069 (cit. on p. 164).

[234] A. León, A. Simon García S., *et al.*, "An advanced search system to manage SARS-CoV-2 and COVID-19 data using a model-driven development approach," *IEEE Access*, vol. 10, pp. 43 528–43 534, 2022, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3169268 (cit. on p. 164).

[235] A. Bernasconi, A. García S., *et al.*, "A comprehensive approach for the conceptual modeling of genomic data," in *Conceptual Modeling - 41st International Conference, ER 2022, India, October 17-20, 2018*, Accepted for publication (cit. on p. 164).

[236] A. Bernasconi, S. Ceri, A. Campi, and M. Masseroli, "Conceptual modeling for genomics: Building an integrated repository of open data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, vol. 10650 LNCS, 2017, pp. 325–339, ISBN: 9783319699035. DOI: 10.1007/978-3-319-69904-2_26 (cit. on p. 164).

[237] S. a. Ceri, "Overview of GeCo: A Project for Exploring and Integrating Signals from the Genome," en, in *Data Analytics and Management in Data Intensive Domains*, L. Kalinichenko *et al.*, Eds., ser. Communications in Computer and Information Science, Cham: Springer International Publishing, 2018, pp. 46–57, ISBN: 978-3-319-96553-6. DOI: 10.1007/978-3-319-96553-6_4 (cit. on p. 164).

[238] D. J. Rigden *et al.*, "The 2022 nucleic acids research database issue and the online molecular biology database collection," *Nucleic Acids Research*, vol. 50, pp. D1–D10, D1 Jan. 7, 2022, ISSN: 0305-1048. DOI: 10.1093/nar/gkab1195 (cit. on p. 165).

[239] K.-P. Koepfli *et al.*, "The genome 10k project: A way forward," *Annu. Rev. Anim. Biosci.*, vol. 3, no. 1, pp. 57–111, 2015 (cit. on pp. 166, 167).

[240] S. A. Pomponi, "The Global Invertebrate Genomics Alliance (GIGA). 2014. Developing Community Resources to Study Diverse Invertebrate Genomes," English, *Journal of Heredity*, vol. 105, no. 1, pp. 1–18, 2014, Publisher: Oxford University Press, ISSN: 0022-1503. DOI: 10.1093/jhered/est084 (cit. on p. 166).

[241] The Darwin Tree of Life Project Consortium, "Sequence locally, think globally: The Darwin Tree of Life Project," en, *Proceedings of the National Academy of Sciences*, vol. 119, no. 4, Jan. 2022, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2115642118 (cit. on p. 167).

[242] H. A. Lewin *et al.*, "Earth BioGenome Project: Sequencing life for the future of life," en, *Proceedings of the National Academy of Sciences*, vol. 115, no. 17, pp. 4325–4333, Apr. 2018, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1720115115 (cit. on p. 167).

[243] A. Rhie *et al.*, "Towards complete and error-free genome assemblies of all vertebrate species," en, *Nature*, vol. 592, no. 7856, pp. 737–746, Apr. 2021, Number: 7856 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-021-03451-0 (cit. on p. 167).

[244] D. P. Genereux *et al.*, "A comparative genomics multitool for scientific discovery and conservation," en, *Nature*, vol. 587, no. 7833, pp. 240–245, Nov. 2020, Number: 7833 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2876-6 (cit. on p. 167).

[245] D. Jebb *et al.*, "Six reference-quality genomes reveal evolution of bat adaptations," en, *Nature*, vol. 583, no. 7817, pp. 578–584, Jul. 2020, Number: 7817 Publisher: Nature

Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2486-3 (cit. on p. 167).

[246] Y. Hu *et al.*, "Cell kinetics of auxin transport and activity in Arabidopsis root growth and skewing," en, *Nature Communications*, vol. 12, no. 1, p. 1657, Mar. 2021, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-021-21802-3 (cit. on p. 168).

Appendix A

The Complete List of Publications

Code	Title	Year	Venue/Journal
	Journals		
J1	Towards the Understanding of the Human Genome: A Holistic Conceptual Modeling Approach	2020	IEEE Access
J2	Using conceptual modeling to improve genome data management	2021	Briefings in Bioinformatics
J3	On how to generalize Species-Specific Conceptual Schemes to generate a Species-Independent Conceptual Schema of the Genome	2021	BMC Bioinformatics
J4	A Conceptual Model-Based Approach to Improve the Representation and Management of Omics Data in Precision Medicine	2021	IEEE Access
J5	An Advanced Search System to Manage SARS-CoV-2 and COVID-19 Data Using a Model-Driven Development Approach	2021	IEEE Access

Table A.1 continues on the next page

Table A.1 continued from previous page

Code	Title	Year	Venue/Journal
J6	The Challenge of Managing the Evolution of Genomics Data Over Time: a Conceptual Model-based approach	2022	BMC Bioinformatics
J7	Integration of clinical and genomic data to enhance precision medicine: a case of study applied to the retina-macula	2022	Software and Systems Modeling
	Conferences		
C1	VarSearch: Annotating Variations using an e-Genomics Framework	2018	International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE)
C2	Towards an Effective and Efficient Management of Genome Data: An Information Systems Engineering Perspective	2019	Forum at the International Conference on Advanced Information Systems Engineering (CAiSE)
C3	Towards the Generation of a Species-Independent Conceptual Schema of the Genome	2020	CMLS Workshop at the International Conference on Conceptual Modeling (ER)
C4	Applying User Centred Design to Improve the Design of Genomic User Interfaces	2021	International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE)
C5	ISGE: A conceptual Model-based Method to correctly manage genome data	2021	Forum at the International Conference on Advanced Information Systems Engineering (CAiSE)
C6	Evolution of an Adaptive Information System for Precision Medicine	2021	Forum at the International Conference on Advanced Information Systems Engineering (CAiSE)
C7	A Model-based Application for the Effective and Efficient Management of Data associated with Retina-Macula Pathology	2021	Exploring Modeling Methods for Systems Analysis and Development (EMMSAD)

Table A.1 continues on the next page

Table A.1 continued from previous page

Code	Title	Year	Venue/Journal
C8	Characterization and Treatment of the Temporal Dimension of Genomic Variations: A Conceptual Model-Based Approach	2021	CMLS Workshop at the International Conference on Conceptual Modeling (ER)
C9	An Ontological Characterization of a Conceptual Model of the Human Genome	2022	Forum at the International Conference on Advanced Information Systems Engineering (CAiSE)
C10	CitrusGenome: A Bioinformatics Tool to Characterize, Visualize, and Explore Large Citrus Variant Datasets	2022	WALS Workshop at the International Conference on Web Engineering (ICWE)
C11	A comprehensive approach for the conceptual modeling of genomic data	2022	International Conference on Conceptual Modeling (ER)
C12	Conceptual Modeling-based Cardiopathies Data Management	2022	CMLS Workshop at the International Conference on Conceptual Modeling (ER)
C13	A Comparative Analysis of the completeness and concordance of data sources with cancer-associated information	2022	CMLS Workshop at the International Conference on Conceptual Modeling (ER)
C14	An Initial Empirical Assessment of an Ontological Model of the Human Genome	2022	EmpER Workshop at the International Conference on Conceptual Modeling (ER)
	Book Chapters		
B1	Guidelines for Designing User Interfaces to Analyze Genetic Data. Case of Study: GenDomus	2018	Evaluation of Novel Approaches to Software Engineering (revisited selected papers)
B2	GenesLove.Me 2.0: Improving the Prioritization of Genetic Variations	2019	Evaluation of Novel Approaches to Software Engineering (revisited selected papers)
B3	CitrusGenome: Applying User Centered Design for Evaluating the Usability of Genomic User Interfaces	2021	Evaluation of Novel Approaches to Software Engineering (revisited selected papers)

Table A.1: Detailed list of Publications.