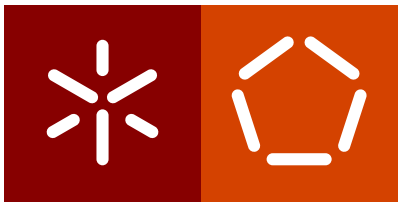
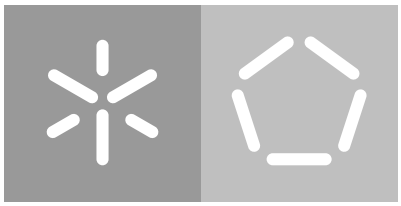


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Bárbara Cláudia Usha Ochs da Fonseca

Genomes Comparisons Through Immune Related Genes In Mammalian Species

June 2020



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Bárbara Cláudia Usha Ochs da Fonseca

Genomes Comparisons Through Immune Related Genes In Mammalian Species

Master dissertation
Master Degree in Computer Science

Dissertation supervised by
Prof. Agostinho Antunes
Prof. Miguel Rocha

June 2020

DIREITOS DO AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

ACKNOWLEDGEMENTS

In the first place, I would like to thank my supervisor, Professor Agostinho Antunes, for this amazing opportunity.

I would also like to acknowledge my co-supervisor, Professor Miguel Rocha.

To Liliana Silva, all the acknowledgments in the world are not enough, as she was the one who put up with all my stress and calmed me down when everything looked like falling apart.

I would like to thank Tito Mendes who made me think deeper about this dissertation.

And to the rest of the people of Evolutionary Genomics and Bioinformatics Group, Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR) and Faculdade de Ciências da Universidade do Porto (FCUP) thank you for being so welcoming.

My family was present during this path, and for that I am extremely grateful.

I would like to thank my friends for the support and affection. They celebrate my achievements as their own.

Finally, I would like to acknowledge my partner in this journey, Inês Fernandes, for the bus trips, lunches and all the adventures in Oporto.

This work was partially supported by the Strategic Funding UID/Multi/04423/2019 through national funds provided by FCT and the European Regional Development Fund (ERDF) in the framework of the program PT2020, by the European Structural and Investment Funds (ESIF) through the Competitiveness and Internationalization Operational Program—COMPETE 2020 and by National Funds through the FCT under the project PTDC/CTA-AMB/31774/2017 (POCI-01-0145-FEDER/031774/2017).

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

Dideoxy sequencing method developed by Sanger was the first sequencing method coming into existence shortly after the report of the DNA structure by Watson and Crick in 1953. DNA sequencing opened up a whole new world in Genomics research and soon whole genomes started to be sequenced.

Among other subjects, comparative evolutionary studies, phylogenetic analyses, mutation frequency assessments, can now be known from such genomic data. For this purpose, researchers' interest relies on having lower cost but highly reliable methods. Currently, there are three generations of sequencing technologies.

Here, we considered two different groups of technologies such as PacBio and older methods (like Illumina and Bac-by-Bac) and used fifty three immune-related genes to assess the best quality of sequencing technologies on eight mammalian genomes (two for human *Homo sapiens* − RBJD01 and DAAB01, two for domestic cat *Felis catus* − AANG04 and ACBE01, two for two for greater horseshoe bat *Rhinolophus ferrumequinum* − RXPC01 and AWHA01 and two for platypus *Ornythorinchus anatinus* − RZJT01 and AAPN01) and an extra genome of the Canadian lynx *Lynx canadensis* for comparability purposes.

The election of the immune related genes relied on the importance of this system, since its function is to protect organisms from external agents which would be seriously threatening in the absence or malfunction of the previously mentioned system. Moreover, immune related genes are often components of large gene-families and show increased levels of genetic variation, which could complicate their proper identification depending on the quality of the genome sequencing.

Quality parameters like number of fragments, integrity of gene and number of artifacts were assessed when screening the selected markers. Although the number of artifacts were not conclusive considering they were underrepresented, the number fragments is lower in NOD-like receptors and Toll-like receptors genes and the integrity is higher in Interferon receptors genes, C-type Lectin receptors genes, NOD-like receptors genes (sub-family P included) and Toll-like receptors genes for the genomes of *Homo sapiens*,*Felis catus*, *Ornythorinchus anatinus* and *Rhinolophus ferrumequinum* sequenced with newly methods.

We conclude that PacBio sequenced genomes showed higher contiguity and better quality overall than the ones sequenced with the older methods.

Keywords: Genome, Third Generation Sequencing, Mammals, Immunity.

RESUMO

O método de sequenciação Dideoxy desenvolvido por Sanger foi o primeiro método de sequenciação que existiu, aparecendo pouco depois da estrutura de DNA ser reportada por Watson e Crick em 1953. A sequenciação de DNA deu origem a novas oportunidades na investigação Genómica e rapidamente genomas completos começarem a ser sequenciados.

Entre outras disciplinas, estudos de comparação evolutiva, análises filogenéticas, avaliação da frequência de mutações podem ser obtidos através de dados genómicos. Para este propósito, é do interesse do investigador fazer uso de métodos de baixo custo, mas fidedignos.

Aqui, consideramos dois grupos diferentes de tecnologias, como PacBio e métodos antigos (como Illumina e Bac-by-Bac), e usámos cinquenta e três genes relacionados com o sistema imunológico para avaliar a qualidade das tecnologias de sequenciamento em oito genomas de mamíferos (dois de humano *Homo sapiens* − RBJD01 e DAAB01, dois de gato doméstico *Felis catus* − AANG04 e ACBE01, dois de morcego-de-ferradura-grande *Rhinolophus ferrumequinum* − RXPC01 e AWHA01 e dois de Ornitorrinco *Ornythorinchus anatinus* − RZJT01 e AAPN01) e um genoma extra de lince canadense *Lynx canadensis* com a finalidade de métrica de comparação.

A eleição dos genes do sistema imunológico baseou-se na importância desse sistema, uma vez que sua função é proteger os organismos de agentes externos que seriam seriamente ameaçadores na ausência ou mau funcionamento do sistema mencionado anteriormente. Ainda, genes relacionados com o sistema imunológico são frequentemente membros de famílias de genes extensos e mostram grandes níveis de variação genética, o que pode complicar sua identificação adequada, dependendo da qualidade da sequenciação do genoma.

Parâmetros de qualidade como número de fragmentos, integridade do gene e número de artefatos foram avaliados ao analisar os marcadores selecionados. Embora o número de artefatos não tenha sido conclusivo, o número de fragmentos é menor em genes das famílias NOD-like receptors e Toll-like receptors e a integridade é maior em genes das famílias Interferon receptors, C-type Lectin receptors, NOD-like receptors (subfamília P incluída) e Toll-like receptors para os genomas de *Homo sapiens*, *Felis catus*, *Ornythorinchus anatinus* e *Rhinolophus ferrumequinum* sequenciados com o método mais recente.

Concluímos que os genomas sequenciados com PacBio apresentaram maior contiguidade e melhor qualidade no geral do que os sequenciados com os métodos mais antigos.

Palavras-chave: Genoma, Sequenciação de Terceira Geração, Mamíferos, Imunidade.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

The amount of available data on biological sequences has greatly expanded recently due to the increase molecular evolution and bioinformatics assessments. Bioinformatics may be defined as the employment of computational tools and analyses' techniques to collect and interpret biological data. By combining disciplines as computer science, biology, physics and mathematics, this specialty became essential to modern biology, medicine, and pharmacy industry (Bayat, 2002).

The main applications of Bioinformatics consist, largely, in computer programs (software) and on the internet platforms. To facilitate the access to biological information, computer networks were developed, and data analyses was eased with the development of useful software. A good example of easy access to information includes biological databases, but because the information is so diverse it was not possible to develop only one database with all the information. Instead, multiple and more specific databases emerged (Benton, 1996).

The genetic sequences are studied by a field designated Comparative Genomics (Krzywinski et al., 2009).

Regarding this field of research, phylogenetic models, ecological interactions, interpretation of phenotypic traits related to the species' ancestry and responses to nucleotide and/or amino acid variation can be attained by studying molecular changes. These may appear in the form of mutations and lead to adaptive significance by promoting changes at biochemical, physiological and morphological levels. To unravel these complex mechanisms linked to organisms' adaptation based on their DNA and proteins changes, computational methods became essential (Baxevanis and Ouellette, 2004).

Molecular evolution can thus be acute for the study of several relevant questions, such as the evolution of sensory/defense mechanisms (Khan et al., 2015), the impact of certain species in an ecosystem (Gomes et al., 2016), or even the origin and expression of human morphological/physiologic developmental disorders related with genetic and environmental diseases (Uhl et al., 2008).

1.1 CONTEXT AND MOTIVATION

The purpose of the immune system is to protect the organism against external agents like, for instance, bacteria and viruses. On a daily basis, nearly every species faces threats by exposure to the previously mentioned agents and, if it were not for their immunological skills, they would be always fallen ill.

Infectious outbreaks are regulated by the genetics of host populations, as well as by the interaction of the genetic structure of parasites, which changes to adapt and escape host defenses. When a disease outbreak is violent, a population of many individuals may be reduced to only a few, and whether these individuals survive because of isolation and lack of exposure or they had resistance genes, the next generation may not reproduce just because of chance, or predator success, or altered sex ratio, etc. Another problem that arises is that, even though species instinctively avoid mating with close relatives, they see them in a situation where they are forced to. (O'Brien and Evermann, 1988)

The numbers are unsure, but many species are currently extinct and several others have small population sizes which put them at risk (Frankham et al., 2002). The immune system evolves through creating diversity, but inbreeding and consequential loss of genetic diversity are inevitable for species in this condition and therefore, the populations' ability to face environmental alterations is weakened. To ensure these species survival, human action towards salvation is now needed (Frankham, 2003).

Thus, the motivation of this dissertation hangs in the importance of expanding the insight on the immune system by inferring the huge genetic diversity of genes involved in immunity. In order to do this, we assess distinct available strategies of genomes sequencing (old strategies like Bac-by-Bac and Illumina vs the third-generation one PacBio) by conducting an intense study of in-depth genome information and analysis with the objective of determining the best approach to be used in further essays.

1.2 MAIN GOALS

Genomics, proteomics and bioinformatics tools can reveal signatures of selection, genetic variability and mutation, and then realize the molecular adaptations of the immune response to infection.

The use of third-generation sequencing technology (e.g. PacBio) that produces several 100 fold better genome assemblies than current second-generation genome sequencing would allow putatively to make much stronger statements on the molecular evolution of large gene families, such as those of the Immunome, and better understand the role of the interplay between single-site mutation and gene copy number variation.

The main goal here is to comparatively assess five immunological genes families (Interferon receptors genes, C-type Lectin receptors genes, NOD-like receptors genes (subfamily P belongs to a different group because it is a large subfamily with the total of fourteen genes) and Toll-like receptors genes from eight mammals' genomes obtained using third-generation sequencing technologies relatively to others obtained with less advanced sequencing methods.

1.3 THESIS ORGANIZATION

This document is organized into six chapters.

In the first chapter, the theme was introduced by giving a general contextualization of the interest of this dissertation as well as the main goals to achieve.

The second chapter will address literature revision. This chapter has 8 sections that harness Immunity (Adaptive and Innate), the Immunome, Immune receptors of interest such as Interferon Receptors, C-type Lectin Receptors, NOD-like receptors, Toll-like receptors, Killer Immunoglobulin Receptors, Sequencing Methods like BAC sequencing, Illumina and PacBio and, last, Tools and Software.

The third chapter is Methods and it will clarify the steps made towards obtaining the results and how these were treated.

The fourth chapter frames the Results and Discussion section that will report the outcomes of this work and its contextualization.

The fifth chapter comprehends the Conclusions. Here, are state the final reviews and further work will be exposed.

The last chapter is Appendix with other tables with results.

2

STATE OF THE ART

2.1 IMMUNITY

2.1.1 *Innate Immunity*

Pattern-recognition receptors (PRRs) recognize preserved and largely invariable molecules like nucleic acids or lipopolysaccharides, which are crucial for microbial physiology. That's on what innate immunity reckons. These essential molecules often go by the name of Pathogen-Associated Molecular Patterns (PAMPs) (Akira et al., 2006; Janeway Jr and Medzhitov, 2002).

When a PAMP is detected, PRR sets in motion diverse mechanisms that trigger the inflammatory and immune responses, and, also, help to assemble an adaptive immune response (Medzhitov, 2007). Examples of this are:

- Toll-Like Receptors (TLRs) are the best-defined class of PRRs. They sense viral nucleic acids and other bacterial products. These can be expressed either on the cell surface or in intracellular compartments (Kawai and Akira, 2006).
- C-type lectin receptors (CLRs), which are expressed at the cell surface, sense sugar motifs of microbial components (Geijtenbeek et al., 2004).
- NOD-like receptors (NLRs) and RIG I-like receptors (RLRs) spot bacteria and viruses that enter the cytoplasm, then promote the production of cytokine production and cell activation (Fritz et al., 2006; Takeuchi and Akira, 2007). Additionally, to operate in circulation, tissue fluids and take part in cell lysis or opsonization of microbes, PRRs like collectins, ficolins and pentraxins are segregated (Medzhitov, 2007).

2.1.2 *Adaptive Immunity*

T cell receptors (TCRs) and B cell receptors (immunoglobulin receptors) mediate adaptive immunity. Variable and constant fragments through the Recombination-Activating gene (RAG) protein-mediated somatic recombination encodes the antigen receptors of T and B

lymphocytes (Cooper and Alder, 2006; Flajnik and Du Pasquier, 2004). Because of this, there is a big range of different antigen receptors, and to increase these diversity mechanisms like non-templated nucleotide addition, gene conversion and somatic hypermutation (in B cells). This concedes great variability in adaptive immune recognition (Schatz et al., 1992).

Lymphocytes that express antigen receptors can be divided into two different types: conventional and innate-like. Related to conventional ones, they are put together randomly and their characteristics are not predetermined (Bendelac et al., 2001). On the other hand, the process of putting together an antigen of innate-like lymphocytes is restricted and their specificities are biased towards a specific group of ligands. Cells with antigens, like dendritic cells, phagocyte microbial antigens into antigenic peptides which meet regular T cells at the cell's top, by means of major histocompatibility complex (MHC) class I and/or class II molecules (McDevitt, 2000).

External agents undergo phagocytosis by cells with antigens and the resulting product is processed into antigenic peptides, which later meet the T cells through Major Histocompatibility Complex (MHC) molecules and, specifically, MHC class Ib which function as PRRs. Conventional B cells bind to an epitope, which is a 3D structure and recognizes it. (McDevitt, 2000)

2.2 IMMUNOME

Similarly to genome, proteome and kinome representing the set of genes, proteins and kinases, respectively, the *immunome* represents the set of genes and proteins involved in immunological mechanisms (Ortutay et al., 2007). This word has also been used to describe the totality of antibodies and antigen receptors (Pederson, 1999).

A total of 847 genes concerning functions such as cell surface recognition, DNA processing, among others were reported in humans (Ortutay et al., 2007).

2.3 TOLL-LIKE RECEPTORS

Toll-like receptors were the first family of PRRs to be discovered when researchers noticed that *Drosophila melanogaster* embryos could not establish a proper dorsal-ventral axis if the gene that encodes to Toll membrane proteins is mutated (Anderson et al., 1985).

Toll protein cytoplasmic domain is homologous to the vertebrate receptor of cytokine IL-1 and by searching for homologous domains to those of Toll and IL-1R in humans, it was discovered a gene that encodes for a similar protein and activates a signaling pathway (Medzhitov et al., 1997). Soon, other relatives to Toll were discovered and they were named Toll-Like Receptors.

TLRs contribute to the normal operation of the immune system in mammals. Research in mice showed that TLR4 mutant gene leads to a defective protein, and, even by being altered in one amino acid only, it no longer recognizes LPS (endotoxin found in cell walls of gram-negative bacteria) and the signaling cascade is compromised (Poltorak et al., 1998).

These membrane proteins can be found either on the plasma membrane or in the membranes of lysosomes or endosomes, as shown in Figure 2. They share a common component called leucine-rich repeats that combined, form a horseshoe-shaped domain. When this domain binds to a PAMP (Figure 3), they are induced to dimerized as a homodimer or as a heterodimer (Jin and Lee, 2008).

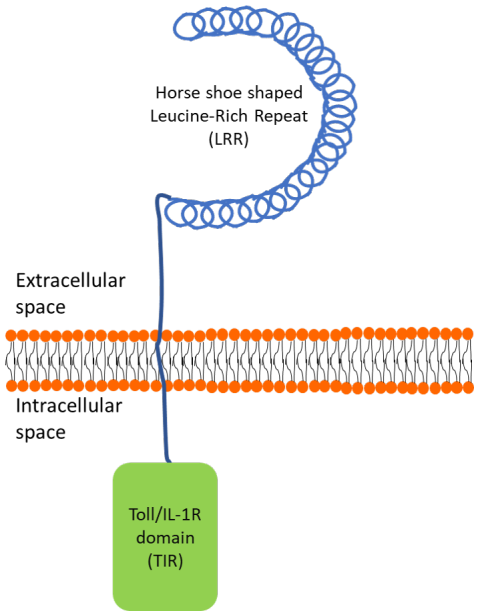


Figure 2: Toll-like receptor (TLR) structure. Structure of a TLR polypeptide chain. The ligand-binding exterior domain of TLRs contains many leucine-rich repeats (LRRs), a transmembranar domain, and an interior Toll/IL-1R (TIR) domain, which interacts with the TIR domains of other members of the TLR signal transduction pathway.

After the PAMP binds to the TLR and dimerization occurred, the signal transduction pathway is determined by the protein adapter. MyD88 (myeloid differentiation factor 88) and TRIF (TIR domain-containing adapter-inducing IFN-b factor) are the two key adapters that are recruited to TLR dimers. Almost every TLR uses MyD88, while TRIF uniquely associates with TLR3 and TLR4 when they are localized in the endosomes (Kindt et al., 2007).

After the dimmer and MyD88 are bind, this structure enrolls with IRAK1 and IRAK4 (IL-1 receptor-associated kinase 1 and 4). IRAK1 phosphorylates itself and TRAF6 (tumor necrosis factor receptor-associated factor 6) activating it, and the latter serves as an organizing center for subsequent signaling components. Proteins TAB1 and TAB2 (TAK1-binding proteins 1 and 2) that come attached to TAK1 (transforming growth factor-β-activated kinase 1) are put

into proximity with IRAK1 that phosphorylates and, consequentially, activates it (Barton and Medzhitov, 2003).

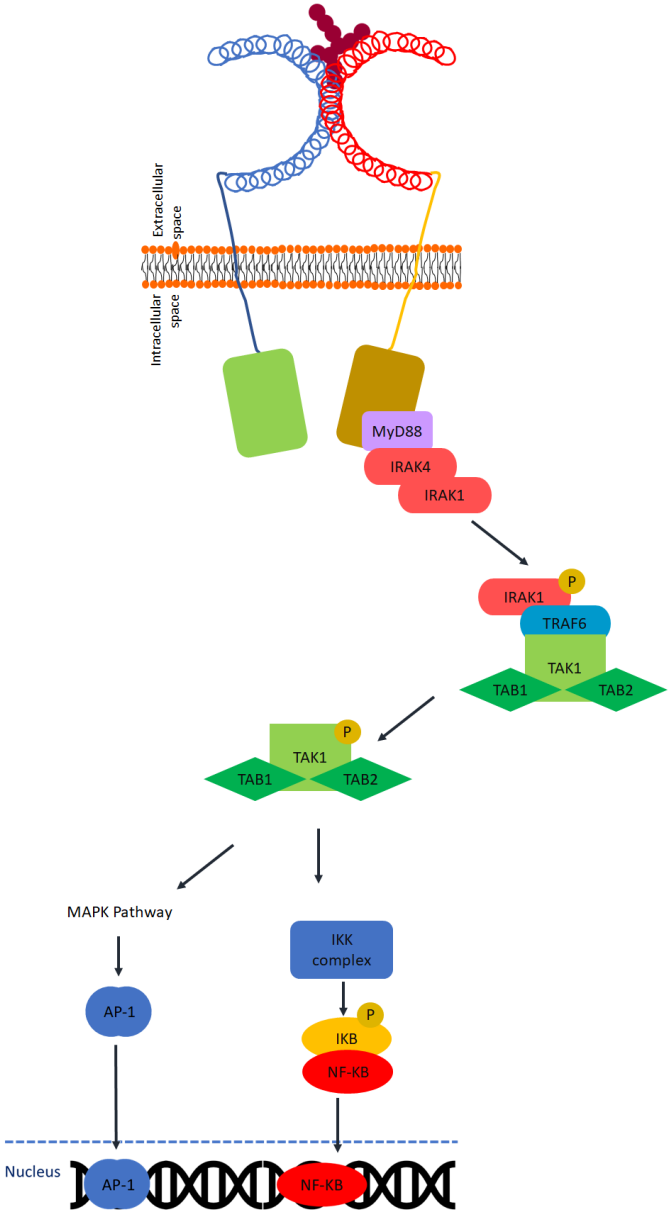


Figure 3: Toll-like receptor 1 and 2 (TLR1 and TLR2) signaling pathway. Signaling pathways downstream of TLR2/1, which binds to a bacterial PAMP, as an example of cell membrane TLRs. TLR undergoes dimerization after being induced by the PAMP. The MyD88 adapter initiates the cascade by enrolling the IRAK1 and IRAK4 kinases. Additional proteins are recruited, including TRAF6 and the TAK1 kinase complex, leading to phosphorylation of the later and activation of MAP kinase pathways, which activate transcription factors such as AP-1, and the IKK complex, leading to the activation of NF-KB.

TAK1 has a double-action in this cascade. On one hand, recruits the IKK (Inhibitor of KB kinase) complex, that is comprised by IKKα, IKKβ and NF-KB Essential Modifier (NEMO), and phosphorylates IKKβ which makes the IKK complex active and free to phosphorylate the Inhibitor of NF-KB. This leads to the release of NF-KB that will enter the nucleus and activate gene transcription. On the other hand, it activates the MAPK signaling pathway which leads to AP-1 dimer formation and this will also activate gene transcription (Takeda and Akira, 2004).

If the TLR dimer binds to TRIF, the cascade proceeds until Interferon Regulatory Factors are activated and induce Interferon α and β genes transcription. This signaling cascade can also lead to the activation of AP-1 or NF-KB (Kindt et al., 2007).

Related to innate and adaptive immune responses, Toll-like receptor family members have been investigated in human disease context. Their function is related to diseases like sepsis, asthma and atherosclerosis (Cook et al., 2004).

2.4 C-TYPE LECTIN RECEPTORS

C-Type Lectin Receptors (CLRs) are located on the cell membrane of a variety of cells related to the immune system, such as monocytes, macrophages, dendritic cells, neutrophils, B cells, and T-cell subsets. These receptors are usually sensitive to carbohydrate components of fungi, mycobacteria, viruses, parasites, and some allergens like peanut and dust mite proteins (Kindt et al., 2007).

As the TLRs, CLRs activate a signaling cascade resulting in the transcription of genes. Taking dectin-1 as an example (Figure 4), this one binds to the PAMP as a dimer and takes place kinase-mediated phosphorylation of one residue of tyrosine in its cytoplasmic domain, more specifically on a structure called half-ITAM (immunoreceptor tyrosine-based activation motif). Next in the chain of events, phospholipase Cδ (PLCδ) is activated, followed by activation of several CARD-containing complexes (Caspase Recruitment Domains-containing complexes). In consequence, by an increase of Calcium concentration inside the cell, these prompt NFAT, NF-KB and AP-1 activation through MAPK pathways.

These transcription factors enter the nucleus and promote gene transcription of proinflammatory cytokines (Gross et al., 2006; Dennehy and Brown, 2007). Dectin-1 also promotes the production of INF-β by activating IRF5 (Kindt et al., 2007).

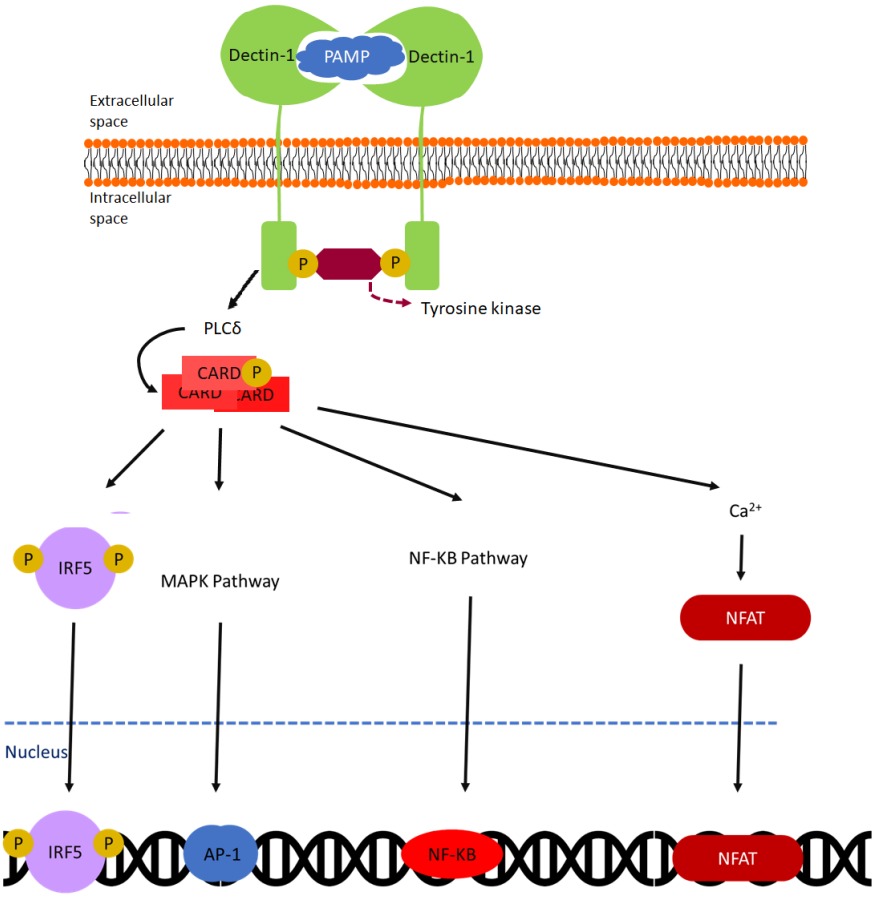


Figure 4: C-type lectin receptor (Dectin-1) signaling pathway. Signaling pathways downstream of the CLR dectin-1. Dectin-1 binds fungal PAMP as a dimer. The tyrosine in the half-ITAM (immunoreceptor tyrosine-based activation motif), located in the cytoplasmic domain of each dectin-1, is phosphorylated, initiating signaling pathways that activate transcription factors NFAT, NF-B, IRF5, and AP-1.

2.5 NOD-LIKE RECEPTORS

Nucleotide Oligomerization Domain/Leucine-rich Repeat-containing receptors are more commonly known as NOD-like receptors (NLRs). These large family of cytosolic proteins activated by intracellular PAMPs are very important to immune and inflammatory responses, but some of them also trigger inflammation, which can be injurious (Inohara and Nuñez, 2001).

Based on their structure, NLRs can be divided into three groups:

- **NLRCs** - They have CARDs;
- **NRLBs** - They have baculovirus inhibitory repeat (BIR) domains;
- **NRLPs** - They have pyrindomains (PYDs).

Despite some NOD receptors are well characterized, like NOD1 and NOD2, for the majority of them the available information is reduced. These two mentioned above, bind to breakdown products of peptidoglycans found in bacterial cell wall, specifically, NOD1 binds to diaminopimelic acid (Chamaillard et al., 2003) and NOD2 to muramyl dipeptides (Girardin et al., 2003).

These receptors associate with endosomal membrane, where their Leucine-Rich Repeats (LRR) bind to PAMPs (Figure 5) and binding the CARD parts of the receptor to RIP2 (Receptor-Interacting Protein kinase 2). Next, TAK1/TAB complex gets attached to RIP2 and activates MAPK pathways and the IKK complex, the latter resulting in NF-KB pathway activation. As seen before, the active transcription factors AP-1 and NF-KB induce the transcription of inflammatory cytokines, antimicrobial and other mediators. Moreover, RIP2 can activate the TRAF3 complex which will lead to phosphorylation of IRF3 and IRF7 that will promote the production of Type I interferons (Windheim et al., 2007).

NOD1 and NOD2 initiate autophagy to eliminate cytosolic bacteria. The bacteria is surrounded by the membrane of endoplasmic reticulum which forms an autophagosome that will merge with lysosomes killing the bacteria (Travassos et al., 2010).

Changing the subject to NLRPs, these contain PYD domains that connect and aggregate with other proteins. When LPS is detected by the cell, a big complex containing NLRs and mature caspase-1 cleaves pro-IL-1β turning it into its mature form, which is the major cytokine produced during innate and inflammatory responses (Martinon et al., 2002). NLRP1, NLRP3, and NLRC4 have been shown to form inflammasomes that cleave pro-IL-1β and pro-IL-18 and turn them into their mature form (Vance, 2015; Chavarría-Smith and Vance, 2015).

Mutations in the NLRP3 gene are associated with auto-inflammatory diseases because they stimulate excessively caspase-1 activity. This inflammasome consists of multiple copies each of NLRP3, the adapter protein ASC (which binds to NLRP3 via homotypic PYD-PYD interactions), and caspase-1 (which binds to ASC via homotypic CARD-CARD interactions), thus forming a large complex (Martinon et al., 2002). The interest in this particular NOD-like receptor doesn't end here. In addition to components from bacteria, fungi and some viruses, NLRP3 can be also activated by DAMPs (Danger-Associated Molecular Patterns) like β-amyloid which is associated with Alzheimer's plaques (Saresella et al., 2016).

2.6 IFN Receptors Family

Interferons (IFNs) are proteins that interact with external agents limiting cell proliferation and with immunomodulating properties (Pestka et al., 1987). They are classified into 3 types (I, II and III) depending on their receptors and response. Type I of IFN include IFN-α and I IFN-β, whereas IFN-γ is a type II and IFN-λ, a type III. (Branca and Baglioni, 1981).

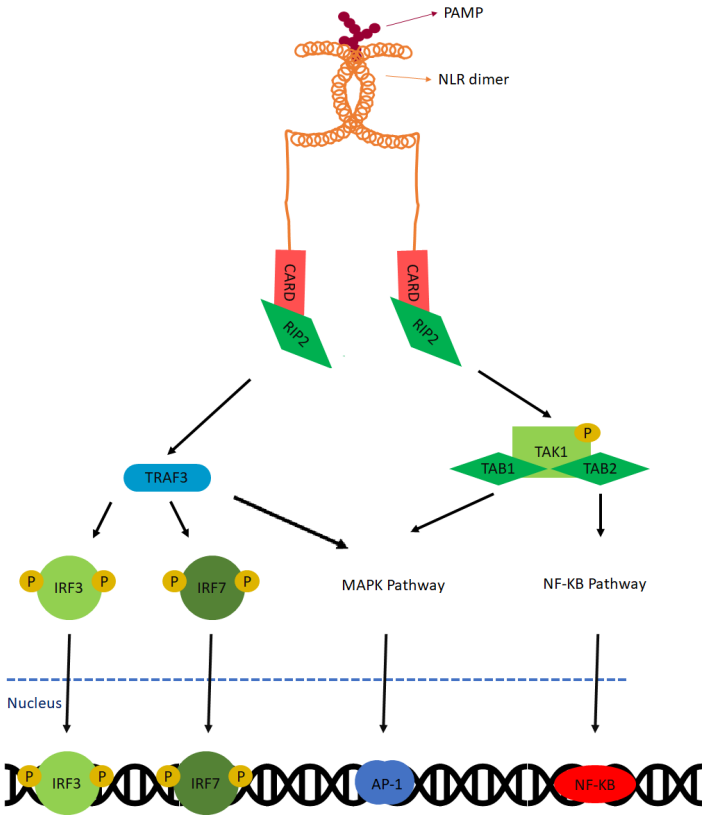


Figure 5: NOD-like receptor 1 (NOD1) signaling pathway. NOD1 associates with endosomes, where it binds to diaminopimelicacid from bacteria. After dimerization NOD1 dimers enroll with RIP2 (receptor-interacting protein kinase 2), which then binds the TAK1/TAB complex, activating MAPK pathways and also the NEMO/IKK complex. The later initiates the NF-KB activation pathway. NOD1 binding of RIP2 can also activate TRAF3, leading to the phosphorylation and activation of IRF3 and IRF7.

IFN I receptor encompasses two sub-units: IFNAR1 - the alpha sub-unit; and IFNAR2 - the beta sub-unit (**Fig.**6). The association of these results in a site with high affinity for IFNs type I (Uzé et al., 1990; Novick et al., 1994).

The signal transduction pathway of IFN-α/β receptor happens when this one binds to a type I IFN. Then, signal transducers and activators of transcription proteins (p84/p91 and p113) suffer Tyrosin phosphorylation and combine with p48 to form the ISGF3 complex. After entering the nucleus, ISGF3 binds to ISRE (cis-acting IFN-stimulated response elements) that exist in IFN-induced genes and initiate their transcription. Tyk2 and JAK1 are the enzymes that phosphorylate the ISGF3 (Müller et al., 1993) (Velazquez et al., 1992). This receptor and Tyr kinase JAK1 are physically connected, which means that signals are directly transduced across the cell membrane (Novick et al., 1994).

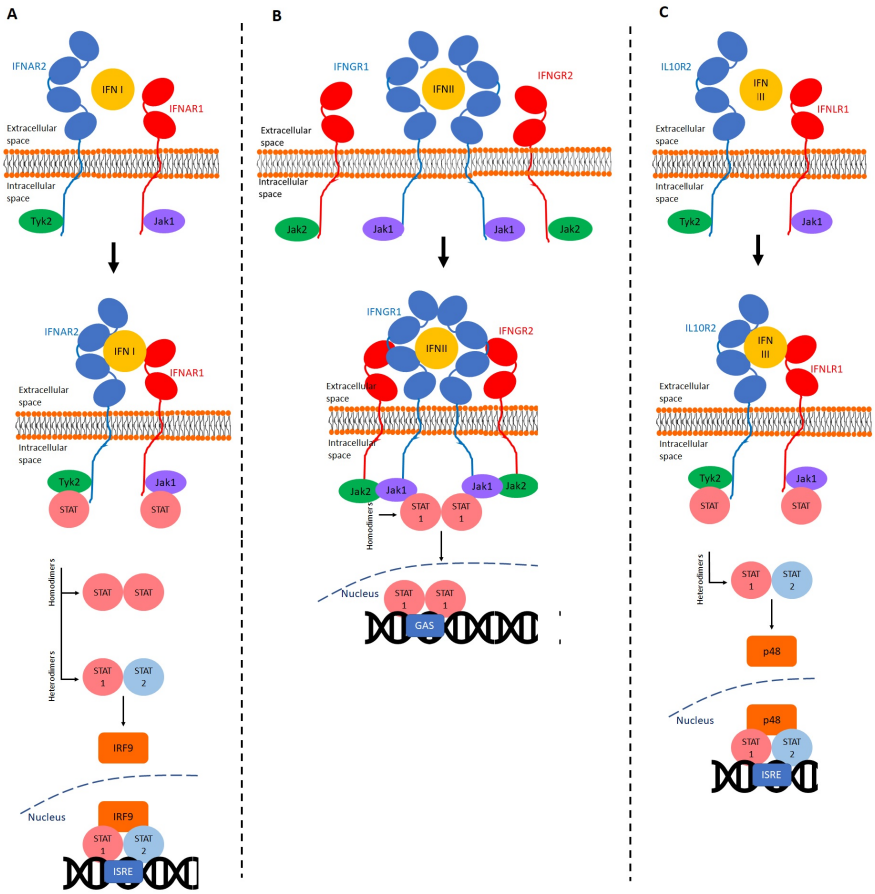


Figure 6: Signaling pathways of IFN receptors. A-α/β receptor; B-γ receptor; C-λ receptor. (Thomas et al., 2011)

IFN-γ receptor comprises two sub-units: the binding sub-unit (IFN-γR1), cell-surface protein encoded on chromosome 6 in humans (Rashidbaigi et al., 1986); and the second chain of the complex receptor (IFN-γR2), also known as Accessory Factor (AF-1) (Soh et al., 1994) located on the human chromosome 21 (Jung et al., 1988).

The cascade of events begins when the IFN-γ binds to IFN-γR1 and then, IFN-γR2 attaches too, forming the receptor complex (Hibino et al., 1992). Jak1 links to IFN-γR1 and Jak2 with IFN-γR2. The 2 chains undergo oligomerization and the kinases (Jak1 and Jak2) get associated with the part of the receptor inside the cell. These two are brought together and activated by phosphorylation. STAT1a enrolment with the receptor complex results in Tyr-701 phosphorylation and, subsequently, in its homodimerization (Shuai et al., 1993). After the STAT1a dimer is formed as a consequence of the IFN-γ binding, it translocates itself to the nucleus, more specifically at the promoter regions of IFN-γ-inducible genes,

where it interacts with the GAS element and the induction of this gene's family begins (Kotenko et al., 1995).

The IFN III receptor (IFNλR) consists of a unique IFN-λR1 chain and a shared IL-10R2 chain. When heterodimerization is induced by IFNλ, the signal transduction pathway of JAK-STAT begins. This comprises the phosphorylation of STAT2 and ISFG3 complex activation, with later up-regulation of MHC class I antigen expression (Kotenko et al., 2003).

2.7 KILLER CELL IMMUNOGLOBULIN-LIKE RECEPTORS

The Killer Cell Immunoglobulin-like Receptors (KIR) are one type of inhibitory Natural Killer Cell receptors (NKRs) expressed at NK cells and a subset of T cells. Usually, they bind to a protein called the Major Histocompatibility Complex (MHC) class I molecules expressed in all cells carrying a nucleus (Marsh et al., 2003).

MHC class I molecules are glycoproteins located at the surface of the cell that function as highly specialized antigen-presenting molecules as they display stable complexes with peptide ligands for further T cell (Janeway Jr et al., 2001).

Consequent to KIR engagement to MHC class I molecules, the activation of NK cells is blocked as well as their functions (Marsh et al., 2003). The expression of KIR in NK cells is stochastic initially, but they undergo an educational process to find the perfect balance between killing dangerous cells to the organism and defend healthy self-cells (Fauriat et al., 2010).

2.8 SEQUENCING METHODS

The First Generation Sequencing (FGS) methods accommodate Maxam and Gilbert's chemical chain termination (Maxam and Gilbert, 1977), Sanger's sequencing (or dideoxy method) (Sanger et al., 1977) and, for commercial use, Applied Biosystems was the first corporation to present ABI Prism 3700. The latter was used in the Human Genome Project (Venter et al., 2001).

The primary form of Next Generation Sequencing (NGS), also known as Second Generation Sequencing (SGS), was GS 20 by 454 Life Sciences, posteriorly acquired by Roche which then gave origin to other platforms as GS FLX titanium, GS FLX Titanium+, GS FLX Titanium XLR700 and GS Junior. Other platforms introduced Genome Analyzer, GA II, HiSeq 2000, HiSeq 100, MiSeq (by Illumina), SOLiD 3, SOLiD4 (by Applied Biosystems) and Polonator G.007 (by Dover and Harvard Med School (Schatz et al., 2010).Illumina and Roche/454 are the most used platforms.

As for the Third Generation Sequencing techniques, they are also known by the name of Next-'Next Generation Sequencing' (Schadt et al., 2010). These technologies are based on

single-molecule sequencing (SMS) generating long sequence reads. They can be divided into three categories (Treffer and Deckert, 2010):

1. Fluorescence-based methods (ex: Single-Molecule Real-Time sequencing - SMRT)
2. Non-fluorescent systems (ex: Nano-edges)
3. Raman-based methods (ex: Surface-Enhanced Raman Spectroscopy - SERS)

BAC-by-BAC sequencing is based on breaking the genome into smaller fragments. These fragments will be incorporated into Bacterial Artificial Chromosome (BAC). Bacteria will multiply, generating a ton of copies. Since the plasmid sequence is known, only the foreign DNA is sequenced. Later, the DNA is mixed with DNA polymerase, primers, free nucleotides and terminator nucleotides labeled with fluorescence. The reaction suffers variations on temperature to separate one chain from another, the primer binds to the DNA, the polymerase binds to the primer and starts sequencing until a terminator base is added. This process is repeated several times to achieve fragments of various lengths. These fragments will go through electrophoresis which will separate the fragments by size. In the end, the terminator base of each fragment will light color depending on if it is an A, T, C or G. By converting this color pattern into letters, the genetic sequence can be obtained. The information sequenced is then overlapped to obtain the correct order of the sequence (Zhang and Wu, 2001).

Illumina's methods also consist of getting random fragments of DNA succeeded by binding of known adapters. Once these adapters are linked, additional motifs, like sequencing primer binding sites, indexes and complementary sequence to the flow cell oligo, are added through reduced cycle amplification. Next, comes the time for amplification. Across the lanes of the flow cell there are two types of oligos. Here, the first type of oligos hybridizes with the DNA fragment on one specific end and, then, a polymerase builds a complementary strand. The double-stranded molecule is therefore denatured and the template fragment is washed away. The new strand bents over and attaches to the second type of oligo forming a bridge structure where next a complementary strand is built by a polymerase molecule. The process is repeated and occurs in several clusters simultaneously, cloning every fragment of the template. Coming next, the double chain bridges are denatured, separating the strands and the reverse strands are discharged. The extension of the primer produces the first read where nucleotides tagged with fluorescence are excited and emit colored light. The sequencing product is washed away, and the same steps are repeated to the reverse strand. The DNA is separated based on the indexes attached at the DNA preparation step and subsequently clustered based on their stretches of base calls. By pairing forward and reverse readings, contiguous sequences can be obtained (Srinivasan and Batra, 2014).

SMRT (Single Molecule Real Time) sequencer was introduced by Pacific Biosciences. It promotes long sequence reads in real-time. A single polymerase molecule is attached to

the bottom of each zero-mode waveguide (ZMW) (Eid et al., 2009). A ZMW is a hole of nanometers of size that prevents the visible laser light from passing in its integrity. The ZMW enables the observation of individual molecules against the free labeled nucleotides maintaining a high signal-to-noise ratio. The illumination of this orifice is through its glass support with a detection volume of 20 zeptoliters (20 x 10^{-21} liters). Nucleotides diffuse at microseconds through the hole, but when the correct nucleotide is attached to the growing chain that takes milliseconds. During the incorporation, the fluorescent label it is excited and, when the cleavage is finished, it diffuses away. The sequential bursts of light are detected and recorded. Along with this sequencing technology, comes other PacBio technique, phospholinked nucleotides. Unlike the other sequencing technologies, these free nucleotides have the fluorophore linked to the terminal phosphate instead of the base. This means that when the polymerase cleaves the phosphate group the complementary DNA chain is completely natural, which enables the exploitation of the inherited properties of the polymerase, such as, high speed, long reed length and high fidelity. A different color can be visualized when the polymerase attaches one of these nucleotides to the DNA strand (Schadt et al., 2010).

2.9 TOOLS AND SOFTWARE USED

The growing information and the development of bioinformatics tools to access and assess information promote new findings and deeper knowledge.

From students, passing through science investigators to health-related workers, the range of people using these assets is incredibly wide. This happens because a significant part of what is created is open source and easy to work with, but it does not mean that handling raw data and use complex software can be performed by all.

In this section, a description on what tools and software were used will be provided.

2.9.1 *Biological Databases and Accessing Genomes*

To conduct this study, it was necessary to resort on Biological Databases (DB). Examples of this include:

- Ensembl (Hunt et al., 2018) that processes genomic data of model organisms and Chordata. In this database, there is the alignment of protein and mRNA sequences to the DNA sequence in order to annotate the genes and transcripts. It has tools such as VEP, BLAST or BLATsequence search and assembly converter. The jobs can be saved as long as the user is logged into their system (Cunningham et al., 2014).

- Biogrid (Oughtred et al., 2018) (Biological General Repository for Interaction Datasets) which is an open-access DB where chemical, genetic and protein interactions are annotated for humans and the principal model organisms. The interface is user-friendly and allows a dynamic interaction network view. The user can apply filters to the genetic or protein data, bioactive compounds or targets. Data can be downloaded without restrictions (Chatr-Aryamontri et al., 2017).

- UniProt (uni, 2016) is a protein knowledge resource created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases. It is open access and Advanced search, BLAST, ClustalO, bulk retrieval/download, ID mapping are some of its tools.

- NCBI (`https://www.ncbi.nlm.nih.gov/`) (Coordinators, 2017) is part of the National Institutes of Health (NIH). It combines several databases. GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) which is a genomic DB of annotated sequences that has no restrictions on the usage of data. It is part of the International Nucleotide Sequence Database Collaboration, where the exchange of genomic information happens on a daily basis. Other databases included are Protein (DB for records of protein sequences), PubMed (DB of biomedical literature), and so on.

2.9.2 *Sequences Extraction and Alignment Tools*

One tool of major importance was MEGA - Molecular Evolutionary Genetics Analysis v5.2 (Tamura et al., 2011). Its use relied on sequences alignment, data view, data exportation in fasta format as also for molecular evolution or phylogenetic analyses. This software includes alignment software such as MUSCLE (MUltiple Sequence Comparison by Log-Expectation) (Edgar, 2004a) (Edgar, 2004b) and ClustalW (Coordinators, 2017).

Another important tool was Exonerate v2.2 software (Slater and Birney, 2005). Exonerate is a tool that enables pairwise sequence comparison by sequence alignment using a variety of alignment models. These models can be either exhaustive dynamic programming or a variety of heuristics. This tool also provides visual information of the alignment (Figure 7) by creating a template of amino acid sequence query, symbols representing how identical are the query and the target (”|”, ”!”, ”:”, ”.” and white space, from the most to the least identical), the target in amino acids and, finally, the target in nucleotides. It also provides a quality metric, the "Raw score", which is a value indicative of the quality of the alignment considering the query coverage, the higher the score, better is the match.

```
C4 Alignment:
------------
        Query: IFNAR2_Homo_sapiens
       Target: ACBE01299515.1 Felis catus breed mixed c474401486.Contig1,
       whole genome shotgun sequence:[revcomp]
        Model: protein2genome:local
    Raw score: 171
  Query range: 11 -> 73
 Target range: 1712 -> 1526

  12 : SerLeuAsnLeuValLeuMetValTyrIleSerLeuValPheGlyIleSerTyrAspSerPr :   32
       !!!|||!:!! !..!!!!!  |||! !.!!|||!!! !!|||  !.!!  !     ! !!
       ThrLeuSerSerThrPheSerValSerPheSerPhePhePhePhePheLeuLeuPheProLe
1712 : ACCTTGAGTTCCACATTCAGCGTTTCTTTTTCTTTTTTTTTTTTTTTCCTCCTGTTTCCTTT : 1652

  33 : oAspTyrThrAspGluSerCysThrPheLysMetSerLeuArgAsnPheArgSerIleLeuS :   53
       !||||!  :!!||||||||||||||||!!.|||:!!!!!!!!|||!:!||||||  !|||||||
       uAspLeuSerAspGluSerCysThrLeuLysValThrPheArgSerPheArgLeuIleLeuS
1651 : AGATTTGTCAGACGAATCTTGCACTTTAAAGGTGACATTTCGCAGTTTCCGGCTTATCTTGT : 1589

  54 : erTrpGluLeuLysAsnHisSerIleValProThrHisTyrThrLeuLeuTyrThrIleMet :   73
       ||||||||||||||||||||||||||!.!|||||||||||||||||! !||||||:!!|||
       erTrpGluLeuLysAsnHisSerIleAlaProThrHisTyrThrLeuTrpTyrThrValMet
1588 : CATGGGAATTAAAAAACCATTCGATTGCACCAACTCACTATACGTTATGGTACACAGTCATG : 1529

cigar: IFNAR2_Homo_sapiens 11 73 . ACBE01299515.1 1712 1526 - 171  M 186
vulgar: IFNAR2_Homo_sapiens 11 73 . ACBE01299515.1 1712 1526 - 171 M 62 186
# --- START OF GFF DUMP ---
```

Figure 7: Example of Exonerate v2.2 (Slater and Birney, 2005) output file.

2.9.3 *Programming Languages and Statistical Analysis*

Processes Optimization

Python (van Rossum, 1995) is a programming language that focuses on code readability, its syntax is concise and relies on libraries, modules and frameworks. When compared to other programming languages, one same program developed previously developed in another language, in Python requires fewer lines of code. This language is used for Web and Internet Development, for Scientific and Numeric computing, for Educational purposes, Software Development and Business Applications. Here, this resource was used to create scripts for process optimization.

Statistical Analysis

R (R Core Team, 2013) is a Free Software for graphical and statistical development. Similar to S language, it was developed by John Chamber and colleagues at Bell Laboratories. Among the main reasons to use R are its publication-quality plots, along with the fact that it provides a wide range of graphical and statistical analyses techniques (linear and nonlinear modelling, classical statistical tests, clustering, classification, ...).

R software effectively handles data and has a collection of intermediate tools for data analysis. Its language includes conditionals, loops, recursive functions defined by the user and input/output facilities, which make it a simple but efficient programming language.

3

METHODS

The question here is to determine whether, PacBio or other technology that emerged before, produces higher quality information, i.e. which one is more reliable. To comparatively assess immunological genes families from mammals' genomes obtained with different methodologies, we went through a four-stage process.

First, we selected the information to work on, the species and genomes and genes' families.

Second, we extracted the previously mentioned data. At UniProt, we obtained the amino acid sequence of the proteins from the selected families of the *Homo sapiens* species. Then, a tblastn was performed for every sequence from UniProt and we saved the coding region. The genomes were downloaded mostly from NCBI and the *Lynx canadensis*' from GitHub.

At the third stage, we produced new data. On MEGA software, we built the queries in amino acids. By that time, we were able to run Exonerate v2.2 software and align those amino acid queries to the selected genomes. Then, the aligned sequences were extracted as well as complementary information provided by Exonerate with a Python developed tool.

Finally, the information on the picked parameters were organized in tables and, then the appropriated statistical analysis were performed using R statistical software. These steps are demonstrated in Figure 8 and will be described in detail bellow.

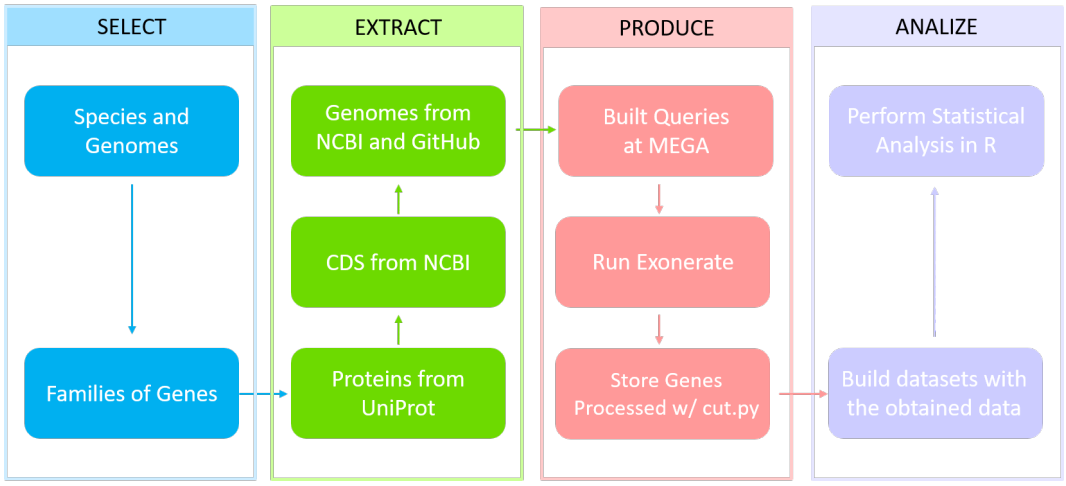


Figure 8: Scheme of the Methodology Process.

1. We selected families of genes involved in immunity due to their importance. In this case, the selected families were INFR, CLR, KIR, NLR and TLR. The NLR family was divided in NOD and NLRP (NOD Like Receptors P). As we worked with mammalian genomes, the model organism for query construction was *Homo sapiens*.

2. From UniProt database, protein sequences of genes from the families mentioned above were downloaded.

3. At NCBI, tblastn (protein to nucleotide) run for each gene retrieved from UniProt. The objective was guarantee that we obtained all available paralogs for each gene family. The choice of the target was the one with the higher "max score", "query coverage" and "par identity" and lower "E-value" the CDS (coding region) of the gene were downloaded in the fasta format.

4. From `https://www.ncbi.nlm.nih.gov/Traces/wgs/` we downloaded genomes of interest. The exception was the genome of *Lynx canadensis* that was retrieved from `https://vgp.github.io/genomeark/Lynx_canadensis/`. In this case, except for the *Lynx Canadensis*, 2 different genomes of the same organism were used. For each species, one of them is sequenced with PacBio and the other one is sequenced by a previous method, as can be seen in Table1.

Table 1: Sequencing method for each genome.

Species	Genome ID	Assembly method
Homo sapiens	RBJD01	PACBIO
	DAAB01	Illumina
Felis catus	AANG04	PACBIO
	ACBE01	Illumina
Ornithorhynchus anatinus	RZJT01	PACBIO
	AAPN01	Shotgun plasmid, fosmid end and BAC
Rhinolophus ferrumrquinum	RXPC01	PACBIO
	AWHA01	Illumina
Lyxn canadensis	primary	PACBIO

5. Later, at MEGA v5.2 (Tamura et al., 2011), the query was constructed with the amino acidic sequence of receptors' genes of *Homo sapiens*.

6. For every genome, Exonerate v2.2 software (Slater and Birney, 2005) was run against the query. The output files contain the regions of the genome were the sequences on the query align. These sequences with good raw scores were extracted to a fasta file with a Python developed tool named *cut.py* (See Appendix– Listing 3.1).

7. Next, these extracted sequences were carefully analyzed to construct a table with parameters that are useful to conclude about the quality of the genome. Those parameters were fragments (in how many fragments were the gene), gene integrity (number of nucleotides extracted/number of nucleotides of the gene in the query), number of ambiguities (number of 'N' in the extracted sequence) and number of artifacts (number of '#' or '*' that may represent a pseudogene).

8. In R software (R Core Team, 2013), each family were analyzed. The data were organized in dataframes. One for Number of Fragments and other for the Integrity of genes of each family. The normality of the data was tested and then the data was tested with the test fitted for the results of the normality. Boxplots were also made to a quick and easy view of data.

Exonerate v2.2 software (Slater and Birney, 2005) was executed through the command `exonerate -model protein2genome -q query.fas -t 1.fsa_nt -showcigar yes -showquerygff yes >2.out`.

The `-model protein2genome` means that we aligned a protein to genomic sequence. The `-q query.fas` was the query file name and `-t 1.fsa_nt` was target file, this meant that every directory had files with these names. `-showcigar yes` is a way of displaying genomic features in 9 fields, these being query identifier, query position at alignment start, query position alignment end, strand of query matched, target identifier, target position at alignment start, target position at alignment end, strand of the target matched and the raw alignment score, as we can see at the bottom of Figure 7. The part `>2.out` is where we put the information into a file, in this case, it was the file "2.out" but the number changes according the directory. For each species genome there are 1 or more fragments, all the fragments were put in a folder number from 1 to the total number of fragments, so the file "2.out" corresponds to the results of the screening where the target file is the second fragment of that genome.

To extract the information from these output files we created a tool named *cut.py* (Listing 3.1) in Python (van Rossum, 1995). An output file consists in a template of aminoacid sequence query, simbols ("|", "!", ":", "." and white space) representing how identic are the query and the target, the target in aminoacids and, finaly, the target in nucleotides. Basically, this is very similar to what is represented in Figure 7 but without the header. Below, is represented the implementation of *cut.py*.

```
import os

path = os.getcwd() #save the directory
for filename in os.listdir(path): #for every file in that directory
    if filename.endswith("cut.txt"):#if it is a file text
        name = filename #save the file name
        f = open(name, 'r', newline = '\n') #open that file
```

```
i = 0
j = 4
inf = []
for line in f: #go through every line of the file
    i += 1
    if i == j : #when it gets to line with the DNA sequence
        inf.append(line) #that line is saved
        j += 5 #jump to the next line containing DNA information
f.close() #close the file

upper=[] #empty list to save all the uppercase letters
for el in inf: #for each element at the inf list (strings of DNA)
    for i in el: #parse through every character
        if i.isupper(): #if it is an uppercase character
            upper.append(i) #it is saved on this list

result = ''.join(upper) #join every element on the upper list on a single string

num = name.count('_') #counts the number of "_" on the origin file
final_name= name.split('_') #variable that holds the name of the original file in
parts
title = ''
for i in range(0, num): #creates a variable to name the new fasta file
    title += final_name[i]
    if i < num-1:
        title+='_'

title1 = title + '.fas' #the name is given by the name of the mother file, without
the "cut.txt" part and adding the fasta file extension

j = open(title1, 'w') #New file with a title created previously
j.write('>' + title + '\n') #the first line is a fasta header characterized by ">"
and followed of the title of the new file
j.write(result) #inserts the string holding the DNA sequence
j.close() #the file is closed
```

Listing 3.1: Implementation of *cut.py*

After we got the genes sequences extracted, we collect the information about the selected parameters and went testing to see if there were significant differences between the two groups. To perform that, the data collected were organized by family, as mentioned before. We use Shapiro-Wilk test to test data for normality of distribution and, due to results, use a non-parametric test, which is compared to the equivalent of the independent t-test, the Mann-Whitney U test. All the tests were performed for a confidence level of 95%.

4

RESULTS AND DISCUSSION

For the present work, it was evaluated whether two different genome sequencing technologies such as PacBio and older methods, like Illumina and Bac-by-Bac, produce sequence quality differences. For that it was performed a screening of immunological markers that were the basis to evaluate some quality parameters like quantity of fragments, wholeness of gene and quantity of artifacts.

The Number of Fragments, Genes Integrity and Artifacts were organized in tables as the one represented in Table 2 for IFNR family. The tables of other families can be consulted in Tables 7 to 11.

The immunological markers we selected for this analysis were found in all species with exception of the KIR family that were only found on the human genome sequenced with PacBio (Table 11). For this reason, these results were not accounted.

Other exception was *Ornythorinchus anatinus*. We verify that this species lacks a lot of the selected immunological markers present in other mammals, like *Homo sapiens*, and we were only able to find genes of CLR soluble and CLR type I subfamilies (only in the genome sequenced with PacBio), 2 genes out of 14 in NLRP subfamily (again, only in the one sequenced PacBio) and NOD and TLR families (in both analyzed genomes). IFNR, CLR type II and KIR were not found. This might be explained by the fact that this species is a very basal mammal. The pool of immunity related molecules reveals major differences, although the major organs of this species are identical to other mammals (Diener and Ealey, 1965). The *O. anatinus* belongs to a taxon called Monotremata that belongs to the subclass Prototheria. Prototherians and Therians divergence occurred around 166 million years ago (Bininda-Emonds et al., 2007).

The phylogenetic distance between *O. anatinus* and other considering mammals in this study is so high that even if this monotreme could present the immunological markers, they might be so distinct that would not align with the query sequence. One example of this unexpected result is that none of the searched KIR genes were found despite it was reported at least 214 genes in Wynne and Tachedjian (2015) (human only present 15 KIR genes (Kelley et al., 2005)).

Despite the limited amount of data retrieved from *O. anatinus* species, most of the genes found were retrieved from PacBio genomes.

Within primates KIR genes have impressively diverge at fast rate with species-specific expansion (Sambrook et al., 2005). For example, gorilla has 11 identified KIR genes, of which only two being human orthologs (Rajalingam et al., 2004); The rhesus macaque solely has five of these genes identified (Hershberger et al., 2001; LaBonte et al., 2001); Chimpanzee has three human orthologs of the seven identified KIR genes (Khakoo et al., 2000). If there are so many differences between primates that are more closely related, it is normal that the differences between humans and the other species are enough to not find genes of this family in the remaining species. Thus, in the human species KIR genes were found just in the PacBio sequenced genome which supports the preposition that this sequencing method is superior in quality.

Table 2: Results of genomes comparisons of Inteferon Receptors. For each genome, it is represented the number of fragments how many different parts the gene is divided; The integrity of the gene ratio between the number of nucleotides extracted from Exonerate and the number of nucleotides from the original query; The number of artifacts either "#" or "*" and they represent insertions and stop codons respectively.

	Genome	**Sequencing**		IFNR				
species	**Prefix**	**Method**		**IFNAR1**	**IFNAR2**	**IFNGR1**	**IFNGR2**	**IFNLR1**
Homo sapiens	RBJD01	PacBio	Number of fragments	1	1	1	1	1
			Integrity of gene	1671/1674	1542/1548	1467/1470	1011/1014	1557/1563
			Artifacts	0	0	0	0	0
	DAAB01	Illumina	Number of fragments	6	6	5	3	6
			Integrity of gene	1470/1674	1449/1548	1386/1470	507/1014	1401/1563
			Artifacts	0	0	0	0	0
Felis catus	AANG04	PacBio	Number of fragments	1	1	1	1	1
			Integrity of gene	1656/1674	1461/1548	1116/1470	981/1014	1539/1563
			Artifacts	0	0	0	0	0
	ACBE01	Illumina	Number of fragments	4	4	5	4	3
			Integrity of gene	1286/1674	1296/1548	1284/1470	762/1014	1347/1563
			Artifacts	0	0	0	0	0
Ornithorhynchus anatinus	RZJT01	PacBio	Number of fragments	0	0	0	0	0
			Integrity of gene	0/1674	0/1548	0/1470	0/1014	0/1563
			Artifacts	-	-	-	-	-
	AAPN01	Shotgun plasmid, fosmid end and BAC	Number of fragments	0	0	0	0	0
			Integrity of gene	0/1674	0/1548	0/1470	0/1014	0/1563
			Artifacts	-	-	-	-	-
Rhinolophus ferrumequinum	RXPC01	PacBio	Number of fragments	1	1	2	1	2
			Integrity of gene	1662/1674	1503/1548	1389/1470	944/1014	1491/1563
			Artifacts	0	0	0	0	0
	AWHA01	Illumina	Number of fragments	1	2	0	1	1
			Integrity of gene	1590/1674	987/1548	0/1470	483/1014	495/1563
			Artifacts	0	0	-	0	0

The number of fragments, gene integrity and artifacts, were subjected to Shapiro-Wilk normality test. The results are shown in Table 3 - 5. For a confidence level of 95%, most of the results have p-values inferior to 0.05 which means that the data do not follow a Gaussian distribution. The exceptions were found at NLRP genes sequenced with the older methods (in genes integrity data frames). In these cases, where only one variable passes the normality test, it was likewise used non-parametric tests in order to ensure the reliability of the results.

The non-parametric test used was the U Test from Mann-Whitney and the results can be consulted in Table 3 - 5. The values signed with ”*” are the ones where the p-value is less than the significance value, and we can reject the null hypothesis and accept the alternative one, where we assume that there are significant differences between the samples. Contrary to the ones not signed with "*" where we fail to reject the null hypotheses and accept that there are no significant changes. Moreover, where the U-Test didn't show significant differences, by looking at the mean and median, we see that those values are always favorable to PacBio group, i.e. in the number of fragments resume table, the mean for the IFNR, CLR and NLRP families is always closer to 1 and the median is exactly 1. When it comes to Old Methods group, although the median for CLR and NLRP is 1, it is 2.5 for IFNR family, and the average of the results ranges from 1.857 to 3.821.

Table 3: Summary results of the fragments number data set. Represented are the mean, median, standard deviation and p-value for Shapiro-Wilk normality test for PacBio and for Old Methods, as well as the p-value for U-test of Mann-Whitney of PacBio against the Old Methods. The p-values in U-Test that are inferior than the significance level are signed with "*".

	Mean		Median		Standard Deviation		Shapiro Wilk Test		U-Test Mann Whitney
	PacBio	Old Methods	PacBio	Old Methods	PacBio	Old Methods	PacBio	Old Methods	p-value
IFNR	0.850	2.550	1.00	2.50	0.5871	2.3050	0.0002	0.0093	0.0544
CLR	0.821	3.821	1.00	1.00	0.0395	0.0863	5.39e-09	0.0005	0.3149
NLRP	0.714	1.857	1.00	1.00	0.5943	3.1009	2.048e-08	2.413e-10	0.7278
NOD	0.969	5.562	1.00	3.00	0.3095	7.9350	3.59e-10	5.83e-08	2.95e-07*
TLR	1.050	2.425	1.00	1.00	0.3889	2.9603	3.4e-12	4.04e-09	0.0006*

Table 4: Summary results of the genes integrity data set. Represented are the mean, median, standard deviation and p-value for Shapiro-Wilk normality test for PacBio and for Old Methods, as well as the p-value for U-test of Mann-Whitney of PacBio against the Old Methods. The p-values in U-Test that are inferior than the significance level are signed with "*".

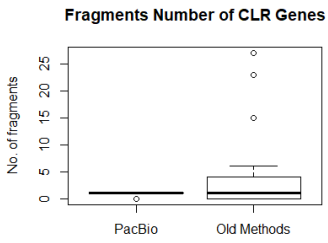
	Mean		Median		Standard Deviation		Shapiro Wilk Test		U-Test Mann Whitney
	PacBio	Old Methods	PacBio	Old Methods	PacBio	Old Methods	PacBio	Old Methods	p-value
IFNR	0.9697	0.7590	0.9846	0.8495	0.0728	0.2004	0.0018	0.0205	1.31e-05*
CLR	0.9824	0.8300	0.9993	0.8635	0.0371	0.1516	2.86e-07	0.0118	5.644e-05*
NLRP	0.9408	0.6745	0.9986	0.6642	0.1057	0.1506	3.057e-08	0.2716	1.272e-08*
NOD	0.9025	0.8017	0.9967	0.8574	0.2160	0.1812	8.064e-09	0.0072	0.0002*
TLR	0.9834	0.9338	0.9969	0.9746	0.0395	0.0863	8.52e-11	2.432e-06	0.0014*

From the previous tables, the information was reorganized to create boxplots (Figures 9 and 10). Based on Figure 9, which correspond to boxplots of the Number of Fragments, we verify that the median of fragments in PacBio sequencing method rounds 1 and it is

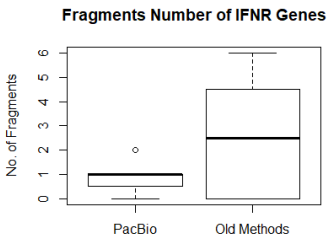
Table 5: Summary results of the artifacts data set. Represented are the mean, median, standard deviation and p-value for Shapiro-Wilk normality test for PacBio and for Old Methods, as well as the p-value for U-test of Mann-Whitney of PacBio against the Old Methods. The p-values in U-Test that are inferior than the significance level are signed with "*". NA - Not Applicable since there are no results.

	Mean		Median		Standard Deviation		Shapiro Wilk Test		U-Test Mann Whitney
	PacBio	Old Methods	PacBio	Old Methods	PacBio	Old Methods	PacBio	Old Methods	p-value
IFNR	NA	NA	NA	NA	NA	NA	NA	NA	NA
CLR	0.565	0.526	0	0	1.9028	1.2188	3.28e-09	5.62e-07	0.5371
NLRP	0.194	0.276	0	0	0.5767	0.5914	3.91e-11	1.41e-08	0.4661
NOD	0.133	0.345	0	0	0.4342	0.8140	1.93e-10	5.05e-09	0.2541
TLR	0.075	0.359	0	0	0.3499	0.8107	3.04e-13	2.94e-10	0.0391*

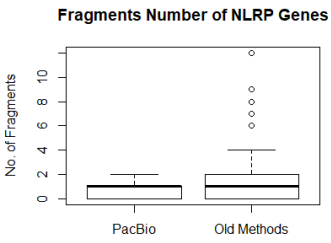
always inferior of the Old Methods. Contrarily in Figure 10, the median of gene integrity in PacBio is globally higher than the older methods. Other aspect that can be observed in these graphics is that the standard deviation is always higher in the older methods.



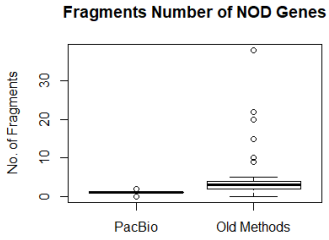
(a) 1a



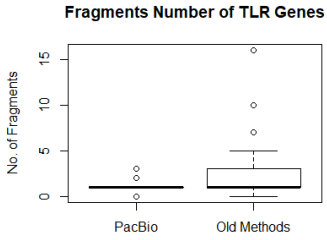
(b) 1b



(c) 1c



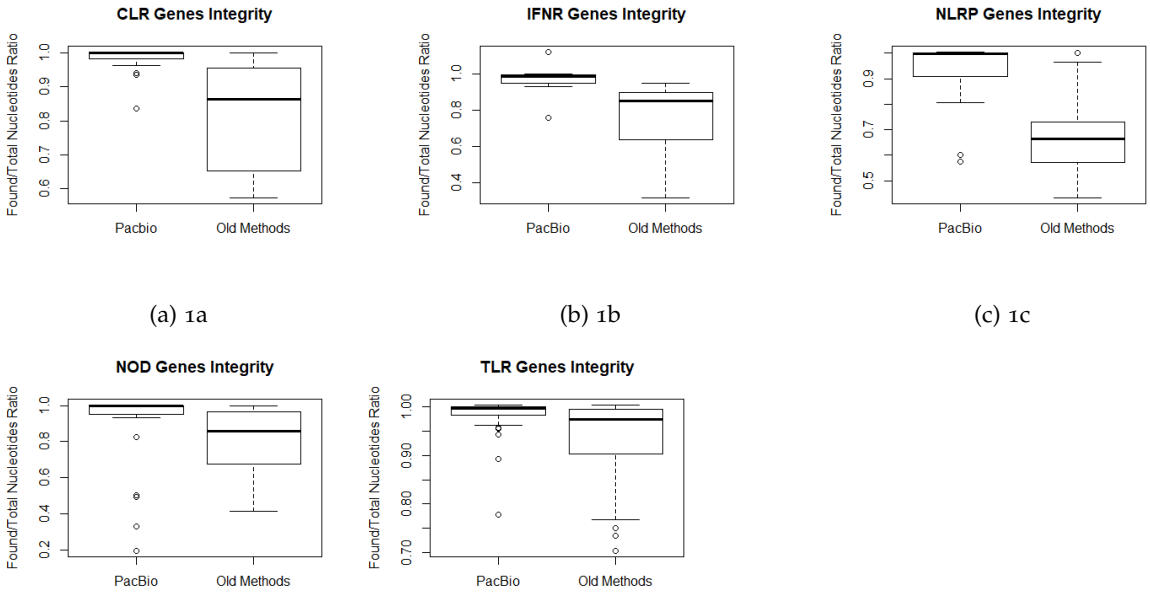
(d) 1d



(e) 1e

Figure 9: Boxplots of the Fragments Number of CLR, IFNR, NLRP, NOD and TLR genes.

The "Number of Fragments" reflects the number of pieces extracted to assemble the gene. The number of fragments is not directly correlated with the gene coverage, but the higher the number of fragments, the higher the possibility of assembly errors as also more chances of loss of nucleotides. Looking at the same gene within the same species, where there are less fragments the coverage is usually higher, as well as the fragments from a lower fragmented gene holds more base pairs when compared to the its more fragmented version.



(a) 1a (b) 1b (c) 1c

(d) 1d (e) 1d

Figure 10: Boxplots of Genes Integrity of CLR, IFNR, NLRP, NOD and TLR genes.

. We detected significant differences at number of fragments between PacBio and older methods only in the NOD and TLR genes families. This might be explained by the size of the sample, as these data sets have more genes as the others. It is easier to obtain lower p-values with bigger size data, which is true for these two. Still on this subject, although these two families shown significant differences, the NOD family is, in fact, the one where the number of fragments is higher in the Old methods group, reaching the 38 fragments in NOD4 gene at the Illumina's genome of *Homo sapiens* species, while in TLR family the maximum number of fragments is 16 in TLR9 on the same genome. The size of genes in those families might be the reason for this discrepancy of NOD genes comparably to other families, once NOD genes size ranges from 2862-5601 base pairs (bp) and, for example TLR genes vary from 2355bp to 3150bp.

The "Integrity of Gene" parameter measures how many nucleotides of the sequence were extracted comparatively to the number of nucleotides exists in the query. The closer the value is to 1, the similar size are shared between extracted sequences and query and, consequently, more integer is the gene. The human genome sequenced with PacBio is the one where the integrity is closer to one. This fact also could be related to the used queries belong to *Homo sapiens* and in remaining species it has the phylogenetic distance effect. In cases where the integrity is higher than one, that is explained by higher size of extracted sequences, comparing with used query, it might occurred nucleotide or codon insertions. Relatively to the integrity between the two sequencing systems in other species, the pattern is sustained,

as the integrity is consistently higher in PacBio. By testing these samples, we get that there are significant differences for all the families, and this is the primary source of evidence that PacBio is a better sequencing technology than the previously existing methods. More than extracting genes in 1 fragment, the important part is to extract the genomic information as complete and intact as possible.

Non-functional copies of genes where the frame is disrupted are called pseudogenes. These can be a pool of diversity (by recombining with their functional paralogous) or even regulate the gene expression. Some genes diversity occurs by conversion of the gene derived from alleles or paralogous genes. This conversion can also occur from pseudogenes, so the sequence, as well as other functions, may be conserved. They could be very conserved and even be active as they can lose some specific functions, retain others or even gain new ones (Ortutay et al., 2007) (Balakirev and Ayala, 2003). Here, they are detected by the emergence of "#" or "***", what we called artifacts, in the middle of the sequence, where the first represents one insertion and the second a stop codon. TLR was the only family where we found significant differences for the different technologies. In IFNR they were not found at all. As the number of artifacts found for the selected immunological markers is reduced, and this is why this is not considered a good parameter for reliable information.

The actual available version of *Lynx canadensis* genome was sequenced with PacBio technology. Here, this genome acts as a comparison metric. It is a very recent genome, that dates of 10/01/2019. Some genes were not found in this species, but the ones we found are discriminated in Table 6. All detected genes were extracted in a single fragment, with NOD4 that is the larger gene in this screening being extracted in two fragments, and the integrity of the genes is also high, generally, with the exception of IFNGR1 which is about half. The missing genes and the lower integrity of IFNGR1 can be explained with loss of information during sequencing, or even loss of certain functions. This comparison metric, produced quality results and supports the premise that the third generation of sequencing methods has higher reliability, which is in accordance to the other information presented here.

Table 6: Results of gene extraction for *Lynx canadensis* genome. For each family of genes, it is represented the number of fragments how many different parts the gene is divided; The integrity of the gene ratio between the number of nucleotides extracted from Exonerate and the number of nucleotides from the original query; The number of artifacts either "#" or "*" and they represent insertions and stop codons respectively.

		CRL soluble	CLR I		CLR II			
		MBL2	**MRC1**	**MRC2**	**CLEC4E**	**CLEC6A**	**CLEC7A**	**CLEC9A**
Lynx canadensis	Number of fragments	1	1	1	1	1	1	1
	Integrity of gene	608/747	4368/4371	4449/4440	618/660	606/630	741/744	726/726
	Artifacts	2	0	0	0	0	3	1

	IFNR				
	IFNAR1	**IFNAR2**	**IFNGR1**	**IFNGR2**	**IFNLR1**
Number of fragments	1	1	1	1	1
Integrity of gene	1653/1674	1500/1548	678/1470	1116/1014	1536/1563
Artifacts	0	0	0	0	0

	NLRP						
	NLRP1	**NLRP2**	**NLRP3**	**NLRP4**	**NLRP5**	**NLRP6**	**NLRP7**
Number of fragments	0	0	1	1	1	1	0
Integrity of gene	-	-	3120/3111	2952/2985	2985/3603	2598/2679	-
Artifacts	-	-	0	0	0	0	-

	NLRP						
	NLRP8	**NLRP9**	**NLRP10**	**NLRP11**	**NLRP12**	**NLRP13**	**NLRP14**
Number of fragments	0	1	1	0	1	1	1
Integrity of gene	-	2976/2976	1854/1968	-	3186/3186	2661/3132	2916/3282
Artifacts	-	0	0	-	0	0	0

	TLR						
	TLR1	**TLR2**	**TLR3**	**TLR4**	**TLR5**	**TLR6**	**TLR7**
Number of fragments	1	1	1	1	0	1	1
Integrity of gene	2352/2361	2354/2355	2715/2715	2478/2520	-	2388/2391	3153/2150
Artifacts	0	1	0	0	-	0	0

	TLR			NOD			
	TLR8	**TLR9**	**TLR10**	**CIITA**	**NOD1**	**NOD2**	**NOD3**
Number of fragments	1	1	1	1	1	1	1
Integrity of gene	3126/3126	3084/3099	2428/2436	3381/3393	2862/2862	3084/3123	3183/3198
Artifacts	0	0	1	0	0	0	0

	NOD			
	NOD4	**NOD5**	**IPAF**	**NAIP**
Number of fragments	2	1	1	1
Integrity of gene	2802/5601	2932/2928	3066/3075	4148/4212
Artifacts	0	13	0	5

5

CONCLUSIONS

5.1 CONCLUSIONS

In this work, we used several immunological markers to compare PacBio sequencing technology against older methods (like Illumina and Bac-by-Bac). The parameters proposed for comparison proposed were the Number of Fragments, Integrity of Gene and Artifacts.

We conclude that in PacBio genomes, the number of fragments is lower and comes with more genetic information than a fragment of a highly fragmented gene and that prevents the loss of information in the assembly stage.

The independent analysis of *Lynx canadensis* genome, which is a recent genome sequenced with PacBio technology, exhibits good results (genes in single fragments and higher integrity). This also enforces the importance of PacBio sequencing technology when compared with older methods.

We also verify that some immunological markers are highly variable across species and the *Ornythorinchus anatinus* lacks a lot of immunological markers present in humans and other mammals. Despite the limited data obtained in this species, it was easily to find information on PacBio sequenced genome.

But the result which supports that PacBio is best preforming of the methods surveyed is without a doubt the gene integrity, that was always closer to one (and, for that reason, better) in this method comparatively to the old methods group.

5.2 PROSPECT FOR FUTURE WORK

Genomics, proteomics and bioinformatics tools can reveal signatures of selection, genetic variability and mutation, and then realize the molecular adaptations. Moreover, regarding these fields of research, phylogenetic models, ecological interactions, interpretation of phenotypic traits related to the species' ancestry and responses to nucleotide and/or amino acid variation can be attained by studying molecular changes which allows the scientific community to unravel complex mechanisms linked to organisms adaptation based on their

DNA and proteins changes. This may provide new solutions for diseases, prevent species extinction and, this way, contribute to a more balanced ecosystem.

There is much other aspects in immune field that can be explored because, for example, humans comprise a total of 847 genes and there are a lot more systems to be studied. It is indispensable to extend the set of genes undergoing the screening process. The same goes to other species like birds where genomes present high G+C content which often leads to genome sequencing mistakes or missing information (Hron et al., 2015). This highlights the need to further test if there are differences between distinct sequencing methods in recording high G+C content sites.

BIBLIOGRAPHY

Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169, 2016.

Shizuo Akira, Satoshi Uematsu, and Osamu Takeuchi. Pathogen recognition and innate immunity. *Cell*, 124(4):783–801, 2006.

Kathryn V Anderson, Gerd Jürgens, and Christiane Nüsslein-Volhard. Establishment of dorsal-ventral polarity in the drosophila embryo: genetic studies on the role of the toll gene product. *Cell*, 42(3):779–789, 1985.

Evgeniy S Balakirev and Francisco J Ayala. Pseudogenes: are they "junk" or functional dna? *Annual review of genetics*, 37(1):123–151, 2003.

Gregory M Barton and Ruslan Medzhitov. Toll-like receptor signaling pathways. *Science*, 300 (5625):1524–1525, 2003.

Andreas D Baxevanis and BF Francis Ouellette. *Bioinformatics: a practical guide to the analysis of genes and proteins*, volume 43. John Wiley & Sons, 2004.

Ardeshir Bayat. Bioinformatics. *Bmj*, 324(7344):1018–1022, 2002.

Albert Bendelac, Marc Bonneville, and John F Kearney. Autoreactivity by design: innate b and t lymphocytes. *Nature Reviews Immunology*, 1(3):177, 2001.

David Benton. Bioinformatics—principles and potential of a new multidisciplinary tool. *Trends in biotechnology*, 14(8):261–272, 1996.

Olaf RP Bininda-Emonds, Marcel Cardillo, Kate E Jones, Ross DE MacPhee, Robin MD Beck, Richard Grenyer, Samantha A Price, Rutger A Vos, John L Gittleman, and Andy Purvis. The delayed rise of present-day mammals. *Nature*, 446(7135):507, 2007.

Andrew A Branca and Corrado Baglioni. Evidence that types i and ii interferons have different receptors. *Nature*, 294(5843):768, 1981.

Mathias Chamaillard, Masahito Hashimoto, Yasuo Horie, Junya Masumoto, Su Qiu, Lisa Saab, Yasunori Ogura, Akiko Kawasaki, Koichi Fukase, Shoichi Kusumoto, et al. An essential role for nod1 in host recognition of bacterial peptidoglycan containing diaminopimelic acid. *Nature immunology*, 4(7):702, 2003.

Andrew Chatr-Aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, et al. The biogrid interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379, 2017.

Joseph Chavarría-Smith and Russell E Vance. The nlrp 1 inflammasomes. *Immunological reviews*, 265(1):22–34, 2015.

Donald N Cook, David S Pisetsky, and David A Schwartz. Toll-like receptors in the pathogenesis of human disease. *Nature immunology*, 5(10):975, 2004.

Max D Cooper and Matthew N Alder. The evolution of adaptive immune systems. *Cell*, 124 (4):815–822, 2006.

NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 45(Database issue):D12, 2017.

Fiona Cunningham, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, et al. Ensembl 2015. *Nucleic acids research*, 43(D1):D662–D669, 2014.

Kevin M Dennehy and Gordon D Brown. The role of the β-glucan receptor dectin-1 in control of fungal infection. *Journal of leukocyte biology*, 82(2):253–258, 2007.

E Diener and EHM Ealey. Immune system in a monotreme: studies on the australian echidna (tachyglossus aculeatus). *Nature*, 208(5014):950, 1965.

Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004a.

Robert C Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1):113, 2004b.

John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

Cyril Fauriat, Martin A Ivarsson, Hans-Gustaf Ljunggren, Karl-Johan Malmberg, and Jakob Michaëlsson. Education of human natural killer cells by activating killer cell immunoglobulin-like receptors. *Blood*, 115(6):1166–1174, 2010.

Martin F Flajnik and Louis Du Pasquier. Evolution of innate and adaptive immunity: can we draw a line? *Trends in immunology*, 25(12):640–644, 2004.

Richard Frankham. Genetics and conservation biology. *Comptes Rendus Biologies*, 326:22–29, 2003.

Richard Frankham, David A Briscoe, and Jonathan D Ballou. *Introduction to conservation genetics*. Cambridge university press, 2002.

Jörg H Fritz, Richard L Ferrero, Dana J Philpott, and Stephen E Girardin. Nod-like proteins in immunity, inflammation and disease. *Nature immunology*, 7(12):1250, 2006.

Teunis BH Geijtenbeek, Sandra J van Vliet, Anneke Engering, Bert A 't Hart, and Yvette van Kooyk. Self-and nonself-recognition by c-type lectins on dendritic cells. *Annu. Rev. Immunol.*, 22:33–54, 2004.

Stephen E Girardin, Ivo G Boneca, Jérôme Viala, Mathias Chamaillard, Agnès Labigne, Gilles Thomas, Dana J Philpott, and Philippe J Sansonetti. Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (mdp) detection. *Journal of Biological Chemistry*, 278(11):8869–8872, 2003.

Cidália Gomes, Ronaldo Sousa, Tito Mendes, Rui Borges, Pedro Vilares, Vitor Vasconcelos, Lúcia Guilhermino, and Agostinho Antunes. Low genetic diversity and high invasion success of corbicula fluminea (bivalvia, corbiculidae)(müller, 1774) in portugal. *PloS one*, 11(7):e0158108, 2016.

Olaf Gross, Andreas Gewies, Katrin Finger, Martin Schäfer, Tim Sparwasser, Christian Peschel, Irmgard Förster, and Jürgen Ruland. Card9 controls a non-tlr signalling pathway for innate anti-fungal immunity. *Nature*, 442(7103):651, 2006.

Karen L Hershberger, Richa Shyam, Ayako Miura, and Norman L Letvin. Diversity of the killer cell ig-like receptors of rhesus monkeys. *The Journal of Immunology*, 166(7):4380–4390, 2001.

Y Hibino, C S Kumar, T M Mariano, D H Lai, and S Pestka. Chimeric interferon-gamma receptors demonstrate that an accessory factor required for activity interacts with the extracellular domain. *Journal of Biological Chemistry*, 267(6):3741–3749, 1992. URL http://www.jbc.org/content/267/6/3741.abstract.

Tomáš Hron, Petr Pajer, Jan Pačes, Petr Bartněk, and Daniel Elleder. Hidden genes in birds. *Genome biology*, 16(1):164, 2015.

Sarah E Hunt, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M Armean, Stephen J Trevanion, Paul Flicek, and Fiona Cunningham. Ensembl variation resources. *Database*, 2018, 11 2018. ISSN 1758-0463. doi: 10.1093/database/bay119. URL https://doi.org/10.1093/database/bay119.

Naohiro Inohara and Gabriel Nuñez. The nod: a signaling module that regulates apoptosis and host defense against pathogens. *Oncogene*, 20(44):6473, 2001.

Charles A Janeway Jr and Ruslan Medzhitov. Innate immune recognition. *Annual review of immunology*, 20(1):197–216, 2002.

Charles A Janeway Jr, Paul Travers, Mark Walport, and Mark J Shlomchik. The major histocompatibility complex and its functions. In *Immunobiology: The Immune System in Health and Disease. 5th edition*. Garland Science, 2001.

Mi Sun Jin and Jie-Oh Lee. Structures of the toll-like receptor family and its ligand complexes. *Immunity*, 29(2):182–191, 2008.

Vincent Jung, Carol Jones, Abbas Rashidbaigi, David D. Geyer, Helvise G. Morse, Rosemary B. Wright, and Sidney Pestka. Chromosome mapping of biological pathways by fluorescence-activated cell sorting and cell fusion: Human interferon gamma receptor as a model system. *Somatic Cell and Molecular Genetics*, 14(6):583–592, Nov 1988. ISSN 1572-9931. doi: 10.1007/BF01535312. URL https://doi.org/10.1007/BF01535312.

Taro Kawai and Shizuo Akira. Innate immune recognition of viral infection. *Nature immunology*, 7(2):131, 2006.

James Kelley, Lutz Walter, and John Trowsdale. Comparative genomics of natural killer cell receptor gene clusters. *PLoS genetics*, 1(2):e27, 2005.

Salim I Khakoo, Raja Rajalingam, Benny P Shum, Kristin Weidenbach, Laura Flodin, David G Muir, Flávio Canavez, Stewart L Cooper, Nicholas M Valiante, Lewis L Lanier, et al. Rapid evolution of nk cell receptor systems demonstrated by comparison of chimpanzees and humans. *Immunity*, 12(6):687–698, 2000.

Imran Khan, Zhikai Yang, Emanuel Maldonado, Cai Li, Guojie Zhang, M Thomas P Gilbert, Erich D Jarvis, Stephen J O'brien, Warren E Johnson, and Agostinho Antunes. Olfactory receptor subgenomes linked with broad ecological adaptations in sauropsida. *Molecular biology and evolution*, 32(11):2832–2843, 2015.

T.J. Kindt, R.A. Goldsby, B.A. Osborne, and J. Kuby. *Kuby Immunology*. W. H. Freeman, 2007. ISBN 9781429202114. URL https://books.google.pt/books?id=oOsFf2WfE5wC.

Sergei V Kotenko, Grant Gallagher, Vitaliy V Baurin, Anita Lewis-Antes, Meiling Shen, Nital K Shah, Jerome A Langer, Faruk Sheikh, Harold Dickensheets, and Raymond P Donnelly. Ifn-λs mediate antiviral protection through a distinct class ii cytokine receptor complex. *Nature immunology*, 4(1):69, 2003.

Serguei V. Kotenko, Lara S. Izotova, Brian P. Pollack, Thomas M. Mariano, Robert J. Donnelly, Geetha Muthukumaran, Jeffry R. Cook, Gianni Garotta, Olli Silvennoinen, James N. Ihle, and Sidney Pestka. Interaction between the components of the interferon receptor complex. *Journal of Biological Chemistry*, 270(36):20915–20921, 1995. doi: 10.1074/jbc.270.36.20915. URL http://www.jbc.org/content/270/36/20915.abstract.

Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

Michelle L LaBonte, Karen L Hershberger, Bette Korber, and Norman L Letvin. The kir and cd94/nkg2 families of molecules in the rhesus monkey. *Immunological reviews*, 183(1): 25–40, 2001.

Steven GE Marsh, Peter Parham, Bo Dupont, Daniel E Geraghty, John Trowsdale, Derek Middleton, Carlos Vilches, Mary Carrington, Campbell Witt, Lisbeth A Guethlein, et al. Killer-cell immunoglobulin-like receptor (kir) nomenclature report, 2002. *Tissue antigens*, 62(1):79–86, 2003.

Fabio Martinon, Kimberly Burns, and Jürg Tschopp. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proil-β. *Molecular cell*, 10(2):417–426, 2002.

Allan M Maxam and Walter Gilbert. A new method for sequencing dna. *Proceedings of the National Academy of Sciences*, 74(2):560–564, 1977.

Hugh O McDevitt. Discovering the role of the major histocompatibility complex in the immune response. *Annual review of immunology*, 18(1):1–17, 2000.

Ruslan Medzhitov. Recognition of microorganisms and activation of the immune response. *Nature*, 449(7164):819, 2007.

Ruslan Medzhitov, Paula Preston-Hurlburt, and Charles A Janeway Jr. A human homologue of the drosophila toll protein signals activation of adaptive immunity. *Nature*, 388(6640): 394, 1997.

M Müller, Carl Laxton, James Briscoe, C Schindler, T Improta, JE Darnell Jr, GR Stark, and IM Kerr. Complementation of a mutant cell line: central role of the 91 kda polypeptide of isgf3 in the interferon-alpha and-gamma signal transduction pathways. *The EMBO journal*, 12(11):4221–4228, 1993.

Daniela Novick, Batya Cohen, and Menachem Rubinstein. The human interferon $\alpha\beta$ receptor: Characterization and molecular cloning. *Cell*, 77(3):391–400, 1994.

Stephen J O'Brien and James F Evermann. Interactive influence of infectious disease and genetic diversity in natural populations. *Trends in Ecology & Evolution*, 3(10):254–259, 1988.

Csaba Ortutay, Markku Siermala, and Mauno Vihinen. Molecular characterization of the immune system: emergence of proteins, processes, and domains. *Immunogenetics*, 59(5): 333–348, 2007.

Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, et al. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541, 2018.

Thoru Pederson. The immunome. *Molecular Immunology*, 36(15):1127 – 1128, 1999. ISSN 0161-5890. doi: https://doi.org/10.1016/S0161-5890(99)00125-X. URL http://www.sciencedirect.com/science/article/pii/S016158909900125X.

Sidney Pestka, Jerome A. Langer, Kathryn C. Zoon, and Charles E. Samuel. Interferons and their actions. *Annual Review of Biochemistry*, 56(1):727–777, 1987. doi: 10.1146/annurev.bi.56.070187.003455. URL https://doi.org/10.1146/annurev.bi.56.070187.003455. PMID: 2441659.

Alexander Poltorak, Xiaolong He, Irina Smirnova, Mu-Ya Liu, Christophe Van Huffel, Xin Du, Dale Birdwell, Erica Alejos, Maria Silva, Chris Galanos, et al. Defective lps signaling in c3h/hej and c57bl/10sccr mice: mutations in tlr4 gene. *Science*, 282(5396):2085–2088, 1998.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

Raja Rajalingam, Peter Parham, and Laurent Abi-Rached. Domain shuffling has been the main mechanism forming new hominoid killer cell ig-like receptors. *The Journal of Immunology*, 172(1):356–369, 2004.

Abbas Rashidbaigi, Jerome A Langer, Vincent Jung, Carol Jones, Helvise G Morse, Jay A Tischfield, John J Trill, Hsiang-Fu Kung, and Sidney Pestka. The gene for the human immune interferon receptor is located on chromosome 6. *Proceedings of the National Academy of Sciences*, 83(2):384–388, 1986.

Jennifer G Sambrook, Arman Bashirova, Sophie Palmer, Sarah Sims, John Trowsdale, Laurent Abi-Rached, Peter Parham, Mary Carrington, and Stephan Beck. Single haplotype analysis demonstrates rapid evolution of the killer immunoglobulin-like receptor (kir) loci in primates. *Genome research*, 15(1):25–35, 2005.

Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

Marina Saresella, Francesca La Rosa, Federica Piancone, Martina Zoppis, Ivana Marventano, Elena Calabrese, Veronica Rainone, Raffaello Nemni, Roberta Mancuso, and Mario Clerici. The nlrp3 and nlrp1 inflammasomes are activated in alzheimer's disease. *Molecular neurodegeneration*, 11(1):23, 2016.

Eric E. Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):R227–R240, 09 2010. ISSN 0964-6906. doi: 10.1093/hmg/ddq416. URL https://doi.org/10.1093/hmg/ddq416.

David G Schatz, Marjorie A Oettinger, and Mark S Schlissel. V (d) j recombination: molecular biology and regulation. *Annual review of immunology*, 10(1):359–383, 1992.

Michael C Schatz, Arthur L Delcher, and Steven L Salzberg. Assembly of large genomes using second-generation sequencing. *Genome research*, 20(9):1165–1173, 2010.

K Shuai, GR Stark, IM Kerr, and JE Darnell. A single phosphotyrosine residue of stat91 required for gene activation by interferon-gamma. *Science*, 261(5129):1744–1746, 1993. ISSN 0036-8075. doi: 10.1126/science.7690989. URL https://science.sciencemag.org/content/261/5129/1744.

Guy St C Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6(1):31, 2005.

Jaemog Soh, Robert J. Donnelly, Serguei Kotenko, Thomas M. Mariano, Jeffry R. Cook, Ning Wang, Stuart Emanuel, Barbara Schwartz, Toru Miki, and Sidney Pestka. Identification and sequence of an accessory factor required for activation of the human interferon Œ receptor. *Cell*, 76(5):793 – 802, 1994. ISSN 0092-8674. doi: https://doi.org/10.1016/0092-8674(94)90354-9. URL http://www.sciencedirect.com/science/article/pii/0092867494903549.

Srilakshmi Srinivasan and Jyotsna Batra. Journal of next generation sequencing & applications. 2014.

Kiyoshi Takeda and Shizuo Akira. Tlr signaling pathways. In *Seminars in immunology*, volume 16, pages 3–9. Elsevier, 2004.

Osamu Takeuchi and Shizuo Akira. Recognition of viruses by innate immunity. *Immunological reviews*, 220(1):214–224, 2007.

Koichiro Tamura, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei, and Sudhir Kumar. Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10):2731–2739, 2011.

Christoph Thomas, Ignacio Moraga, Doron Levin, Peter O Krutzik, Yulia Podoplelova, Angelica Trejo, Choongho Lee, Ganit Yarden, Susan E Vleck, Jeffrey S Glenn, et al. Structural linkage between ligand discrimination and receptor activation by type i interferons. *Cell*, 146(4):621–632, 2011.

Leonardo H Travassos, Leticia AM Carneiro, Mahendrasingh Ramjeet, Seamus Hussey, Yun-Gi Kim, João G Magalhães, Linda Yuan, Fraser Soares, Evelyn Chea, Lionel Le Bourhis, et al. Nod1 and nod2 direct autophagy by recruiting atg16l1 to the plasma membrane at the site of bacterial entry. *Nature immunology*, 11(1):55, 2010.

Regina Treffer and Volker Deckert. Recent advances in single-molecule sequencing. *Current opinion in biotechnology*, 21(1):4–11, 2010.

Elizabeth W Uhl, Marcus Martin, James K Coleman, and Janet K Yamamoto. Advances in fiv vaccine technology. *Veterinary immunology and immunopathology*, 123(1-2):65–80, 2008.

Gilles Uzé, Georges Lutfalla, and Ion Gresser. Genetic transfer of a functional human interferon α receptor into mouse cells: Cloning and expression of its c-dna. *Cell*, 60(2): 225–234, 1990.

G. van Rossum. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.

Russell E Vance. The naip/nlrc4 inflammasomes. *Current opinion in immunology*, 32:84–89, 2015.

Laura Velazquez, Marc Fellous, George R Stark, and Sandra Pellegrini. A protein tyrosine kinase in the interferon $\alpha\beta$ signaling pathway. *Cell*, 70(2):313–322, 1992.

J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, and et. al Smith. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. ISSN 0036-8075. doi: 10.1126/science.1058040. URL https://science.sciencemag.org/content/291/5507/1304.

Mark Windheim, Christine Lang, Mark Peggie, Lorna A Plater, and Philip Cohen. Molecular mechanisms involved in the regulation of cytokine production by muramyl dipeptide. *Biochemical Journal*, 404(2):179–190, 2007.

James W Wynne and Mary Tachedjian. Bat genomics. in. *Bats and Viruses*, pages 315–326, 2015.

Hong-Bin Zhang and Chengcang Wu. Bac as tools for genome sequencing. *Plant Physiology and Biochemistry*, 39(3):195 – 209, 2001. ISSN 0981-9428. doi: https://doi.org/10.1016/S0981-9428(00)01236-5. URL http://www.sciencedirect.com/science/article/pii/S0981942800012365. Plant genomics.

A

APPENDIX

A.1 GENOMES COMPARISONS TABLES

Table 7: Results of genomes comparisons of C-type Lectin Receptors. For each genome, it is represented the number of fragments how many different parts the gene is divided; The integrity of the gene ratio between the number of nucleotides extracted from Exonerate and the number of nucleotides from the original query; The number of artifacts either "#" or "*" and they represent insertions and stop codons respectively.

	Genome	Sequencing		CLR soluble	CLR I		CLR II			
species	Prefix	Method		MBL2	MRC1	MRC2	CLEC4E	CLEC6A	CLEC7A	CLEC9A
Homo sapiens	RBJD01	PacBio	Number of fragments	1	1	1	1	1	1	1
			Integrity of gene	747/747	4371/4371	4440/4440	660/660	630/630	744/744	726/726
			Artifacts	0	0	0	0	0	0	0
	DAAB01	Illumina	Number of fragments	4	23	27	4	0	5	5
			Integrity of gene	684/747	3687/4371	2850/4440	612/660	-	744/744	726/726
			Artifacts	0	5	1	0	-	0	0
Felis catus	AANG04	PacBio	Number of fragments	1	1	1	1	1	1	1
			Integrity of gene	729/747	4368/4371	4449/4440	618/660	606/630	741/744	726/726
			Artifacts	1	0	0	0	0	0	1
	ACBE01	Illumina	Number of fragments	2	15	6	1	2	4	2
			Integrity of gene	645/747	3748/4371	4356/4440	617/660	399/660	483/744	417/726
			Artifacts	0	2	0	0	0	0	1
Ornithorhynchus anatinus	RZJT01	PacBio	Number of fragments	1	1	1	0	0	0	0
			Integrity of gene	624/747	4314/4371	4430/4440	-	-	-	-
			Artifacts	2	0	9	-	-	-	-
	AAPN01	Shotgun plasmid, fosmid end and BAC	Number of fragments	2	0	0	0	0	0	0
			Integrity of gene	477/747	-	-	-	-	-	-
			Artifacts	1	-	-	-	-	-	-
Rhinolophus ferrumequinum	RXPC01	PacBio	Number of fragments	1	1	1	1	0	1	1
			Integrity of gene	744/747	4311/4371	4440/4440	621/660	-	729/744	726/726
			Artifacts	0	0	0	0	-	0	0
	AWHA01	Illumina	Number of fragments	1	0	1	1	0	1	1
			Integrity of gene	744/747	-	4329/4440	588/660	-	489/744	594/726
			Artifacts	0	-	0	0	-	0	0

Table 8: Results of genomes comparisons of NOD-like Receptors P. For each genome, it is represented the number of fragments how many different parts the gene is divided; The integrity of the gene ratio between the number of nucleotides extracted from Exonerate and the number of nucleotides from the original query; The number of artifacts either "#" or "*" and they represent insertions and stop codons respectively.

	Genome	Sequencing		NLRP													
species	Prefix	Method		NLRP1	NLRP2	NLRP3	NLRP4	NLRP5	NLRP6	NLRP7	NLRP8	NLRP9	NLRP10	NLRP11	NLRP12	NLRP13	NLRP14
Homo sapiens	RBJD01	PacBio	Number of fragments	1	1	1	1	1	1	1	1	1	1	1	1	1	1
			Integrity of gene	4422/4422	3189/3189	3111/3111	2985/2985	3603/3603	2679/2679	2943/2943	3144/3147	2976/2976	1968/1968	3102/3102	3186/3186	3132/3132	3282/3282
			Artifacts	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	DAAB01	Illumina	Number of fragments	12	0	4	7	8	0	0	9	12	2	2	6	9	7
			Integrity of gene	3669/4422	-	2163/3111	1871/2985	2061/3603	-	-	3036/3147	1842/2976	1968/1968	2004/3102	1929/3186	2790/3132	2616/3282
			Artifacts	0	-	1	0	0	-	-	0	0	0	0	0	0	1
Felis catus	AANG04	PacBio	Number of fragments	0	0	1	1	1	1	0	1	1	1	0	1	1	1
			Integrity of gene	-	-	3120/3111	2952/2985	2989/3603	2628/2679	-	1893/3147	2976/2976	1854/1968	-	3187/3186	2661/3132	2916/3282
			Artifacts	-	-	0	0	1	0	-	0	0	0	-	1	0	0
	ACBE01	Illumina	Number of fragments	0	0	1	1	0	0	0	1	1	1	0	1	1	1
			Integrity of gene	-	-	1640/3111	1296/2985	-	-	-	1506/3147	1448/2976	1584/1968	-	1704/3186	2082/3132	1572/3282
			Artifacts	-	-	2	0	-	-	-	1	2	0	-	0	0	0
Ornithorhynchus anatinus	RZJT01	PacBio	Number of fragments	0	0	0	0	0	1	0	0	0	0	0	2	0	0
			Integrity of gene	-	-	-	-	-	1542/2679	-	-	-	-	-	3093/3186	-	-
			Artifacts	-	-	-	-	-	3	-	-	-	-	-	0	-	-
	AAPN01	Shotgun plasmid, fosmid end and BAC	Number of fragments	0	0	0	0	0	0	0	0	0	0	0	0	0	0
			Integrity of gene	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artifacts	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Rhinolophus ferrumequinum	RXPC01	PacBio	Number of fragments	0	2	1	1	1	2	0	0	1	2	0	1	1	1
			Integrity of gene	-	2574/3189	3105/3111	2949/2985	2988/3603	2550/2679	-	-	2988/2976	1824/1968	-	2817/3186	2664/3132	3276/3282
			Artifacts	-	1	0	0	0	0	-	-	0	0	-	0	1	0
	AWHA01	Illumina	Number of fragments	0	0	1	1	1	1	0	0	1	1	0	1	1	1
			Integrity of gene	-	-	3120/3111	2952/2985	2985/3603	2598/2679	-	-	2976/2976	1854/1968	-	3186/3186	2661/3132	2916/3282
			Artifacts	-	-	0	0	0	0	-	-	0	0	-	0	0	0

Table 9: Results of genomes comparisons of Toll-like Receptors. For each genome, it is represented the number of fragments how many different parts the gene is divided; The integrity of the gene ratio between the number of nucleotides extracted from Exonerate and the number of nucleotides from the original query; The number of artifacts either "#" or "*" and they represent insertions and stop codons respectively.

	Genome	Sequencing		TLR									
species	Prefix	Method		TLR1	TLR2	TLR3	TLR4	TLR5	TLR6	TLR7	TLR8	TLR9	TLR10
Homo sapiens	RBJD01	PacBio	Number of fragments	1	1	1	1	1	1	1	1	1	1
			Integrity of gene	2362/2361	2355/2355	2715/2715	2250/2520	2577/2577	2391/2391	3150/3150	3123/3126	3099/3099	2436/2436
			Artifacts	1	0	0	0	0	0	0	0	0	0
	DAAB01	Illumina	Number of fragments	3	5	5	3	1	5	10	7	16	4
			Integrity of gene	2211/2361	2292/2355	2715/2715	2508/2520	2535/2577	1683/2391	2910/3150	3063/3126	2274/3099	2403/2436
			Artifacts	0	0	0	0	0	0	0	0	0	0
Felis catus	AANG04	PacBio	Number of fragments	1	1	1	1	1	1	1	1	1	1
			Integrity of gene	2352/2361	2355/2355	2715/2715	2478/2520	2574/2577	2388/2391	3153/3150	2430/3126	3084/3099	2430/2436
			Artifacts	0	0	0	0	0	0	0	0	0	0
	ACBE01	Illumina	Number of fragments	2	3	3	1	1	2	1	2	1	2
			Integrity of gene	2263/2361	1767/2355	2203/2715	2220/2520	2575/2577	2224/2391	3152/3150	2737/3126	3084/3099	2308/2436
			Artifacts	0	3	2	0	1	1	2	1	0	1
Ornithorhynchus anatinus	RZJT01	PacBio	Number of fragments	0	2	1	1	1	1	1	1	3	1
			Integrity of gene	2301/2361	2268/2355	2670/2715	2379/2520	2565/2577	2334/2391	3162/3150	3105/3126	2963/3099	2331/2436
			Artifacts	0	0	0	0	0	0	0	0	2	0
	AAPN01	Shotgun plasmid, fosmid end and BAC	Number of fragments	0	2	1	1	1	1	1	1	0	1
			Integrity of gene	2301/2361	2268/2355	2310/2715	2379/2520	2565/2577	2334/2391	3162/3150	3105/3126	-	2328/2436
			Artifacts	0	0	0	0	0	0	0	0	-	3
Rhinolophus ferrumequinum	RXPC01	PacBio	Number of fragments	1	1	1	1	1	1	1	1	1	1
			Integrity of gene	2352/2361	2358/2355	2676/2715	2481/2520	2565/2577	2373/2391	3150/3150	3123/3126	3087/3099	2436/2436
			Artifacts	0	0	0	0	0	0	0	0	0	0
	AWHA01	Illumina	Number of fragments	1	1	1	1	1	2	1	1	1	1
			Integrity of gene	2328/2361	2358/2355	2082/2715	2112/2520	2565/2577	1938/2391	3150/3150	3123/3126	3099/3099	2436/2436
			Artifacts	0	0	0	0	0	0	0	0	0	0

Table 10: Results of genomes comparisons of NOD-like Receptors. For each genome, it is represented the number of fragments how many different parts the gene is divided; The integrity of the gene ratio between the number of nucleotides extracted from Exonerate and the number of nucleotides from the original query; The number of artifacts either "#" or "*" and they represent insertions and stop codons respectively.

species	Genome Prefix	Sequencing Method		NOD							
				CIITA	NOD1	NOD2	NOD3	NOD4	NOD5	IPAF	NAIP
Homo sapiens	RBJD01	PacBio	Number of fragments	1	1	1	1	1	1	1	1
			Integrity of gene	3393/3393	2862/2862	3123/3123	3198/3198	5601/5601	2928/2928	3075/3075	3486/4212
			Artifacts	0	0	0	0	0	0	1	0
	DAAB01	Illumina	Number of fragments	22	15	20	4	38	10	4	2
			Integrity of gene	2556/3393	2211/2862	2904/3123	1323/3198	5253/5601	2832/2928	2985/3075	1809/4212
			Artifacts	0	0	0	0	0	0	0	0
Felis catus	AANG04	PacBio	Number of fragments	1	1	1	1	1	1	1	1
			Integrity of gene	3381/3393	2862/2862	3084/3123	3183/3198	2823/5601	2790/2928	3066/3075	4206/4212
			Artifacts	0	0	0	0	0	0	0	0
	ACBE01	Illumina	Number of fragments	3	4	2	3	9	2	4	4
			Integrity of gene	2032/3393	2785/2862	2169/3123	3134/3198	4524/5601	1803/2928	2961/3075	3130/4212
			Artifacts	0	1	1	3	1	0	0	0
Ornithorhynchus anatinus	RZJT01	PacBio	Number of fragments	2	1	1	1	1	1	0	0
			Integrity of gene	1680/3393	2853/2862	2907/3123	3096/3198	1095/5601	2775/2928	-	-
			Artifacts	0	0	0	0	0	0	-	-
	AAPN01	Shotgun plasmid, fosmid end and BAC	Number of fragments	3	1	1	4	5	3	0	0
			Integrity of gene	1971/3393	2853/2862	2907/3123	2760/3198	2763/5601	2401/2928	-	-
			Artifacts	0	0	0	0	0	1	-	-
Rhinolophus ferrumequinum	RXPC01	PacBio	Number of fragments	1	1	1	1	1	1	1	1
			Integrity of gene	3402/3393	2862/2862	3054/3123	3198/3198	1851/5601	2860/2928	3069/3075	4194/4212
			Artifacts	0	0	0	0	1	2	0	0
	AWHA01	Illumina	Number of fragments	2	1	3	1	3	0	2	3
			Integrity of gene	2292/3393	2454/2862	3045/3123	3198/3198	3606/5601	-	3069/3075	3639/4212
			Artifacts	0	0	0	0	3	-	0	0

Table 11: Results of genomes comparisons of Killer Immunoglobulin Receptors. For each genome, it is represented the number of fragments how many different parts the gene is divided; The integrity of the gene ratio between the number of nucleotides extracted from Exonerate and the number of nucleotides from the original query; The number of artifacts either "#" or "*" and they represent insertions and stop codons respectively.

	Genome	Sequencing		KIR2DL						KIR3DL		
species	Prefix	Method		KIR2DL1	KIR2DL2	KIR2DL3	KIR2DL4	KIR2DL5A	KIR2DL5B	KIR3DL1	KIR3DL2	KIR3DL3
Homo sapiens	RBJD01	PacBio	Number of fragments	1	1	1	1	1	0	1	1	1
			Integrity of gene	1047/1047	873/1047	822/1026	1134/1134	1127/1128	-	1335/1335	1368/1368	1233/1233
			Artifacts	0	0	0	2	0	-	0	0	0
	DAAB01	Illumina	Number of fragments	0	0	0	0	0	0	0	0	0
			Integrity of gene	-	-	-	-	-	-	-	-	-
			Artifacts	-	-	-	-	-	-	-	-	-
Felis catus	AANG04	PacBio	Number of fragments	0	0	0	0	0	0	0	0	0
			Integrity of gene	-	-	-	-	-	-	-	-	-
			Artifacts	-	-	-	-	-	-	-	-	-
	ACBE01	Illumina	Number of fragments	0	0	0	0	0	0	0	0	0
			Integrity of gene	-	-	-	-	-	-	-	-	-
			Artifacts	-	-	-	-	-	-	-	-	-
Ornithorhynchus anatinus	RZJT01	PacBio	Number of fragments	0	0	0	0	0	0	0	0	0
			Integrity of gene	-	-	-	-	-	-	-	-	-
			Artifacts	-	-	-	-	-	-	-	-	-
	AAPN01	Shotgun plasmid, fosmid end and BAC	Number of fragments	0	0	0	0	0	0	0	0	0
			Integrity of gene	-	-	-	-	-	-	-	-	-
			Artifacts	-	-	-	-	-	-	-	-	-
Rhinolophus ferrumequinum	RXPC01	PacBio	Number of fragments	0	0	0	0	0	0	0	0	0
			Integrity of gene	-	-	-	-	-	-	-	-	-
			Artifacts	-	-	-	-	-	-	-	-	-
	AWHA01	Illumina	Number of fragments	0	0	0	0	0	0	0	0	0
			Integrity of gene	-	-	-	-	-	-	-	-	-
			Artifacts	-	-	-	-	-	-	-	-	-