Novel genetic mechanisms in autoinflammatory diseases

Contribution of somatic and germline variation

Manuel Solís Moruno

DOCTORAL THESIS UPF / 2021

Thesis supervisors

Dr. Ferran Casals López

Dr. Tomàs Marquès Bonet

Experimental and Health Sciences Department

upf. Universitat Pompeu Fabra Barcelona

La pluma es lengua del alma: cuales fueren los conceptos que en ella se engendraren, tales serán sus escritos.

– Don Quijote.

Acknowledgments

Empiezo a escribir estas líneas el 11 de enero, mucho antes siquiera de abordar el manuscrito de la tesis. Pero es que tenía ganas. Durante estas últimas vacaciones de Navidad me ha dado tiempo de pensar en mis más de cinco años en Barcelona, en quién influyó previamente y en quién lo ha hecho durante.

Antes de meterme en faena con el grueso de esta sección es de justicia que le agradezca a Tomàs el haberme abierto las puertas de su laboratorio, así como su pasión por la ciencia y su afán por hacer cosas grandes y aportar algo más que un granito de arena a la comunidad. También por las charlas de videojuegos y literatura fantástica, claro. Y gracias, por supuesto, a Ferran (anteriormente conocido como Ferrán). Mil gracias. Eres el mejor director de tesis que habría podido pedir, alguien en quien me gusta verme reflejado y con quien creo que tengo bastante afinidad. He agradecido muchísimo tanto que seas un enorme científico como que, además, seas alguien calmado, reflexivo y, por encima de todo, buena persona. Yo de (no tan) mayor quiero ser como tú. Y también quiero dar las gracias a nuestros colaboradores del Hospital Clínic, Juantxo y Anna. Ha sido una experiencia muy bonita y enriquecedora compartir proyectos con vosotros.

Y me gustaría ahora sí empezar como tiene que ser, por el principio: la familia. Así que quiero agradecer a mis padres su incuestionable fe y confianza en mí y, a mi hermana, su incomprensible admiración y casi fascinación por este su humilde hermano mayor. Habéis conseguido que me sienta hasta casi casi especial e inteligente. A mis abuelos, por todo su cariño y amor –especial mención a los quesos que me has comprado todo este tiempo, Nano, desde que te dije que no me imaginaba lo caro que era hasta que salí de casa–. También a mis primos, por ser mis anclas y el mejor abrazo posible a las raíces. Vosotros sois esas patadas a un balón en la plaza, esas inocentes y despreocupadas horas jugando al Golden Sun y al Pokémon Oro y Plata en el patio, esos espadazos con palos de madera en la azotea de vuestra tía Ana y, más recientemente, esos no tan inocentes ron con naranja y whisky con Seven-Up. Y siempre seréis esa bandera de Andalucía en medio de la avenida.

También quería mencionar a un puñado de grandísimos profesores, que ni siquiera leerán estas páginas, pero cuyo nombre merece un rinconcito entre ellas porque, de una forma u otra, marcaron mi camino. Profesores de los Salesianos como don Juan (probablemente el mejor y el más c***** de todos, estoy seguro de que él mismo aceptaría con gusto el adjetivo), don José Luis, don Fernando y don Óscar. Profesores de la universidad como Carmen Garnacho, Javier Vitorica (la mejor clase de toda la carrera fue, sin duda, la de la ATP sintasa), José López Barneo, Rafa Chacón, Ricardo Pardal, Sebastián Chávez, Fran Romero Campero, Antonio Núñez… La investigación en Sevilla está en buenísimas manos. No me quiero imaginar lo que podríais hacer con una inversión potente.

Por otro lado, he pensado mucho en mi primer laboratorio, aquel de enfermedades infecciosas en el IBiS en el que hice las prácticas de la carrera y el TFG. Así que muchísimas gracias, Javi, por aceptarme en él y por hacerme sentir verdaderamente valorado, aunque en aquel momento sólo fuese un chaval que estaba terminando Biomedicina, un experimento quizá demasiado nuevo. Gracias a Ana, Pablo y Josan por hacerme ver cómo funcionaba realmente un laboratorio. Vosotros me enseñasteis qué cosas eras importantes y qué no dentro y fuera del mismo. Y gracias a Carmen, mi compañera favorita de trayectos y charlas en el lab y por la que, te lo confieso en estas líneas, tengo especial debilidad.

Por supuesto, también tengo muchísimo que agradecer a todos mis compañeros de estos años. A Clàudia, per ser la meva primera referència al lab, i també la millor referència catalana possible, a pesar de ser yo demasiado simpático para ella (imagínense, *un sevillano y una de Vic…* El chiste se escribe solo). A Marc, una de las mentes más brillantes que he conocido, quien me ayudó muchísimo los primeros meses y de quien seguí aprendiendo todo el tiempo, también fuera de la ciencia. Tus charlas siempre han sido fuente de inspiración. A Irene, mi soplo de viento zaragozano, que me hacía ver que no estaba loco entre tanto catalán, pues entre nosotros teníamos muchísimo en común. Por eso y, por supuesto, por las somáticas. Parte de esta tesis también se apellida Lobón. A Aitor, compañero de piso durante casi cuatro años, experto en CNVs, estadística, trilingüe hasta doler, friki de los antiguos, biblioteca cinéfila y maestro en no saber mucho del tema sabiendo muchísimo del tema. A Jéssica, miembro original del zulo, casi una madre en el lab cuando llegué y feliz madre a tiempo completo ahora mismo. A Raquel, que siempre me cayó especialmente bien, por su durísimo trabajo y dedicación. A veces demasiado duro. A Lukas, por su conocimiento en materia técnica y por lo fácil y hasta divertido que fue publicar el *paper* del cromosoma 1 con él. A Guillem, por sus primeras y necesarias enseñanzas y por ese humor tan parecido al mío.

Y a los que llegaron después que yo. A Esther, por ser una auténtica máquina en el laboratorio. A David, quien de verdad creo que sabe mucho de absolutamente todo. A Marina, por tener un gusto exquisito a la hora de hacer cualquier cosa que tenga que ver con el diseño gráfico. A Luis, con el que congenié inmediatamente en ese café que duró un pelín más de lo estrictamente necesario en horario laboral, con quien he pasado horas hablando de videojuegos y quien me ha ayudado en todo lo posible con los *long reads* y mis preguntas de transcriptómica. A Paula, porque su confianza en sí misma, su determinación y lo claras que tenía las cosas desde el principio me dejaron impresionado y me encantaron. Heredera de Raquel también en lo de trabajar duro. Y bueno, tenía que dejar para el final a mi compañera de fatigas y aquella cuyo hombro ha estado al lado del mío durante años (o su espalda, lo mismo da). Gracias, Laura, por todo; por tus asertivos y acertados comentarios siempre que te los pedía, por tus consejos y ayuda, por tu forma de trabajar, por aguantarme cada vez que me daba la vuelta para enseñarte cualquier tontería. Para el recuerdo queda aquel congreso de la ESHG en Milán, en el que me lo pasé genial y del que disfruté especialmente gracias a ti. Y también aquella mítica noche de fiesta entre tanto científico. Moltíssimes gràcies.

I would also like to mention some other important people from whom I have also learnt a lot. So thank you, Martin and Sojung, for all your knowledge. Thank you, Tom and Jonas, for being so nice guys and bring something different to the lab. Gracias, Fátima, por tu refrescante presencia galleguísima.

Gracias también a los miembros del servei de Genòmica, Núria, Roger y Raquel, por toda la ayuda que me han prestado estos años dentro del laboratorio, en el que siempre he sido como un elefante en una cacharrería.

También ha sido muy importante el grupo de Fisio, que me hizo sentir uno más. Así que, gracias, de verdad, a Julia, Natalia, Sonia, Víctor, Pol, Iván y Albert.

Gracias también al gueto de andaluces por traer todos sus acentos a Barcelona. Gracias a Dani, Amaranta, Jose, Jose y Helena por esas tardes de birritas, o de vermús precio guiri, que me insuflaban años de vida.

Y a mis amigos. A los más viejos como Migue y Fran (o Kiko, o Enrique), que siguen ahí desde hace una veintena de años y que, sospecho y espero, seguirán muchos más. Sois la amistad en pasado, presente y futuro. También a Jose Miguel, que además me sacó de fiesta alguna que otra vez los primeros meses por aquí, los más duros. Os quiero, chavales.

Y, por supuesto, a los de la carrera, con El Bar Coyote al completo a la cabeza. Gracias a Ismael, quien huyó a tiempo y se metió en el embolao de estudiar

Medicina, primer ejemplo a seguir que tuve en clase en cuanto el tío alzó la voz en aquella presentación. Ya te he dicho que sin ti la ciencia se pierde algo muy gordo, pero también me alegro de que supieras decir que no cuando viste el percal. Gracias a Zarza por sus pésimos chistes con los que, evidentemente, me río hasta que me duele la boca. Y por esos apuntes maravillosos. Y por esos días interminables de estudio en la facultad. Y por los picos con alioli. Gracias a Andrea, último y gran fichaje para el grupo, especialmente por esas sesiones desinteresadas de deporte en pleno encierro. Gracias a Jesús por ejemplificar a la perfección lo que debe ser un buen científico y hacer siempre las preguntas adecuadas. Y por ser una increíble persona. Gracias al Kokaso por ser TAN buena gente, con el que además compartí máster, otro que supo huir a tiempo y ahora no puede parar de comprarse casas. Gracias a Carmen por su fuerte carácter y, a la vez, por su jovialidad, sus sonrisas y por estar siempre dispuesta a ayudar. Gracias a Susana, porque esos "¡buenos díaaaasss!" me daban tanto coraje a la vez me gustaban, pues eran lo primero que hacía que me activara medianamente en clase. Créeme, guardo con mucho cariño el recuerdo de esas mañanas en la Sopeña. Gracias a Pepi, faro de Chipiona y tilde de la palabra corazón. Sois todos unos seres de luz y os quiero muchísimo. Vuestros éxitos son mis alegrías.

Gracias también a todas las personas con las que he vivido estos años desde que me fui de Sevilla. Gracias de nuevo a Carmen y Susana por aguantarme cuando era un polluelo recién salido del nido. Tenéis el cielo ganao y lo sabéis. Gracias al Pepaso, al que ya le tenía un cariño enorme que se multiplicó en Carrer de la Mineria. Gracias, de nuevo, a Aitor, por ser el mejor compañero de piso de la historia de los compañeros de piso. Gracias a Dani, que fue una gratísima y bienvenida sorpresa. Ya saben, algo podrá estar bueno… Pero nunca será la tortilla de Dani. Gracias a Marta por las risas de nunca saber si dormiría en el piso (spoiler: no, hoy tampoco). Gracias a Víctor por ser igual de fan que yo de Miyazaki e igual de intenso que yo con el fútbol. Y gracias al buen Luckaso, que siempre tendrá un sitio especial en mi corazón, por compartir sus últimos años de amor gatuno con nosotros.

También quería agradecerle a un par de mentes preclaras el haberme llenado de horas de ocio estos años y evitar que me volviera tarumba del todo. Gracias al bueno de Hidetaka Miyazaki por crear Bloodborne y Dark Souls, y a Brandon Sanderson, quien escribe más rápido de lo que cualquier mortal puede leer por, sobre todo, esa primera trilogía de Nacidos de la Bruma. Y ya que me pongo, a todas las personas que hay detrás de The Office y Los Soprano, que hicieron mucho más llevadero un pandémico 2020.

Como bien sabéis muchos, estos años en Barcelona no han sido sencillos, pero gracias a las personas que nombro aquí he podido no sólo sobrevivir a ellos sino llegar a disfrutarlos. Pero sin duda, la principal culpable de lo bueno que he vivido todo este tiempo has sido tú, Marina. Gracias por este maravilloso viaje, por hacerme mejor persona día tras día, por animarme, por quererme, por saber siempre qué decir, por sacarme a viajar y por cuidarme. Espero de corazón haber estado a la altura de todos vosotros, pero especialmente de ti. Gracias. De verdad, muchísimas gracias.

Y, por último, quiero darle las gracias a todas las personas que me han ayudado con las revisiones del manuscrito. Habéis hecho que la calidad del texto sea muy superior a lo que era en un principio.

Madre mía, me he quedao a gusto, eh.

Abstract

Autoinflammatory diseases are caused by an antigen independent hyperactivation of inflammatory pathways. The first genetic defect linked to an autoinflammatory disease was described almost 25 years ago and, since then, the list has been expanded to more than 40 different disorders and causal genes. In this work, we have applied massive parallel sequencing to explore the genetic mechanisms implicated in these diseases. In addition to the detection of germline genetic variants, we have demonstrated that massive parallel sequencing is suitable to detect and characterize the frequency distribution in different cell populations of somatic causal variants, as well as derived some general recommendations for the detection of this type of variation. We have also developed a novel approach based on chromosome isolation by flow cytometry to characterize a large structural variant in a family with an autoinflammatory disease. Thus, we propose that these alternative genetic models, which have remained largely underexplored because of technical and analytical limitations, may explain an important fraction of the undiagnosed patients of rare disorders, and must be considered in future experimental and study designs.

Resumen

Las enfermedades autoinflamatorias son causadas por una hiperactivación de las vías inflamatorias independiente de antígeno. El primer defecto genético asociado a una enfermedad autoinflamatoria fue descrito hace casi 25 años y, desde entonces, la lista se ha expandido hasta más de 40 enfermedades distintas y genes causales. En este trabajo, hemos aplicado secuenciación masiva paralela para explorar los mecanismos genéticos implicados en estas enfermedades. Además de la detección de variantes genéticas germinales, hemos demostrado que la secuenciación masiva paralela es adecuada para detectar y caracterizar la distribución de frecuencias en diferentes poblaciones celulares de variantes somáticas causales, así como sugerir ciertas recomendaciones generales para la detección de este tipo de variación. También hemos desarrollado una nueva aproximación basada en aislamiento de cromosomas mediante citometría de flujo para caracterizar una variante estructural grande en una familia con una enfermedad autoinflamatoria. Así pues, proponemos que estos modelos genéticos alternativos, que han permanecido ampliamente inexplorados debido a limitaciones técnicas y analíticas, podrían explicar una fracción importante de los pacientes de enfermedades raras sin diagnóstico, y deben ser considerados en futuros diseños de estudio y experimentales.

Preface

Massive parallel sequencing technologies have rapidly evolved from its commercialization, 15 years ago. The scientific and the clinical communities have generated an incredible amount of data using them to study genetic variation in humans and other organisms.

Autoinflammatory diseases are the main focus of this thesis. In this work, they were explored by whole exome and whole genome sequencing, both with Illumina and Oxford Nanopore Technologies data, with the aim to discover new somatic and germline pathogenic genetic variants.

In the Introduction section, human genetic variation, along with the different types of variants, is presented in the first place. After that, massive parallel sequencing technologies, with methodological aspects of the variant discovery processes, are detailed. In the second block of the Introduction, an overview of the immune system is given, followed by an explanation of inflammation in physiological conditions. Then, autoinflammatory diseases are introduced, with a description of the pathogenic mechanisms that underlie these disorders. To close the circle and conclude the Introduction, the genetic variants already discovered for autoinflammatory diseases are enumerated.

Results are divided into four chapters. In Chapter 1, we explore the suitability of whole exome sequencing for the discovery of somatic pathogenic variants, in blood samples, from individuals with autoinflammatory disorders and other primary immunodeficiencies. In Chapter 2, we describe two new variants related to the deficiency of IL-1 receptor antagonist. In Chapter 3, we perform a proof of concept of the flow sorting enrichment technique of chromosome 1 and its sequencing with an Oxford Nanopore MinION device. This technique allows us to, in Chapter 4, describe a novel structural variant related to cryopyrin associated periodic syndrome in a particular family.

Finally, these results are discussed in the context of other studies of the field, also with a look into the limitations of the employed techniques and into the future of genomic studies.

List of abbreviations

AD: autosomal dominant.

AR: autosomal recessive.

bp: base pair.

CAPS: cryopyrin associated periodic syndromes.

CH: clonal haematopoiesis.

CLP: common lymphoid progenitor.

CMP: common myeloid progenitor.

CNV: copy number variant.

DIRA: deficiency of IL-1 receptor antagonist.

FMF: familial Mediterranean fever.

GATK: genome analysis toolkit.

HSC: haematopoietic stem cell.

MAF: minor allele frequency.

MPS: massive parallel sequencing.

NGS: next generation sequencing.

ONT: Oxford Nanopore Technologies.

PID: primary immunodeficiency.

PRR: pattern recognition receptor.

SNV: single nucleotide variant.

SV: structural variant.

TCR: T cell receptor.

VAF: variant allele frequency.

WES: whole exome sequencing.

WGS: whole genome sequencing.

Table of contents

Introduction

Ençiende lô faroliyô de tû oxo prinçeçitâ.

–Antonio Martínez Ares. Comparsa El Perro Andalú (2018).

1. Genetic variation

1.1. Human genetic variation

The human reference genome

The first draft of the human genome was published 20 years ago, in 2001. Two different articles were published in Nature (International Human Genome Sequencing Consortium 2001) and Science (Venter et al. 2001) at the same time since, respectively, a public and a private initiative pursued the same objective. In the first case, a big consortium called the Human Genome Project was created and funded with public resources. They partitioned the human genome into pieces and used thousands (> 20,000) of bacterial artificial chromosomes to physically map and sequence it. In the second case, the private company Celera Genomics, run by Craig Venter, used whole genome shotgun sequencing to accomplish the same goal. Two years after the initial release, the consolidated version of the human reference genome was published (Collins et al. 2003).

Efforts to complete and refine the human genome have continued (International Human Genome Sequencing Consortium 2004) and are still ongoing nowadays. At the moment of writing this thesis, the latest version of the human reference genome is the GRCh38.p13 (Genome Reference Consortium Human Build 38 patch release 13), submitted by the Genome Reference Consortium in February 2019 (it can be found on https://www.ncbi.nlm.nih.gov/grc/human). The first version of the complete human genome has been published, as a preprint, days before the deposit of this thesis (Nurk et al. 2021). This reference, called T2T-CHM13, includes the sequences of all autosomes and chromosome X with no gaps.

The technology used by the Human Genome Project to sequence the human genome was the chain termination sequencing or Sanger sequencing (Sanger, Nicklen, and Coulson 1977). It is part of what has been called first generation sequencing technologies. The second generation is characterized by producing short reads, typically around 150 bp, with high accuracy and throughput; Illumina being the company that has dominated this market. These second generation technologies are usually called next generation sequencing (NGS), although the term "next", 15 years after their commercialization, is considered obsolete. Nowadays, they are commonly called massive parallel sequencing (MPS) technologies, along with the third generation. In the third generation we mainly find two technologies: Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). They are characterized by the production of longer reads, typically 1-200 kb (and up to 2 Mb) (Logsdon, Vollger, and Eichler 2020), although at the cost of reduced base quality and increased price.

All these technologies have been extensively compared and reviewed elsewhere (Heather and Chain 2016; Kchouk, Gibrat, and Elloumi 2017; Kumar, Cowley, and Davis 2019; Pareek, Smoczynski, and Tretyn 2011). In particular, NGS methods were reviewed ten years after their initial commercialization (Goodwin, McPherson, and McCombie 2016).

Thanks to the advent of sequencing technologies, obtaining the whole genome sequencing (WGS) data of one individual is considerably cheaper and more accessible than what it used to be. It is estimated that the cost of the first human genome draft was ~$300 million (although the project was funded with $3 billion), whereas today the cost of one WGS is less than $1,000 (Figure 1). In certain cases, we could just be interested in sequencing the coding part of the genome or a fraction of it, so whole exome sequencing (WES) or targeted sequencing, respectively, are an even more affordable option. As a curiosity, the first individual to have the whole genome sequenced was, precisely, Craig Venter's (Levy et al. 2007).

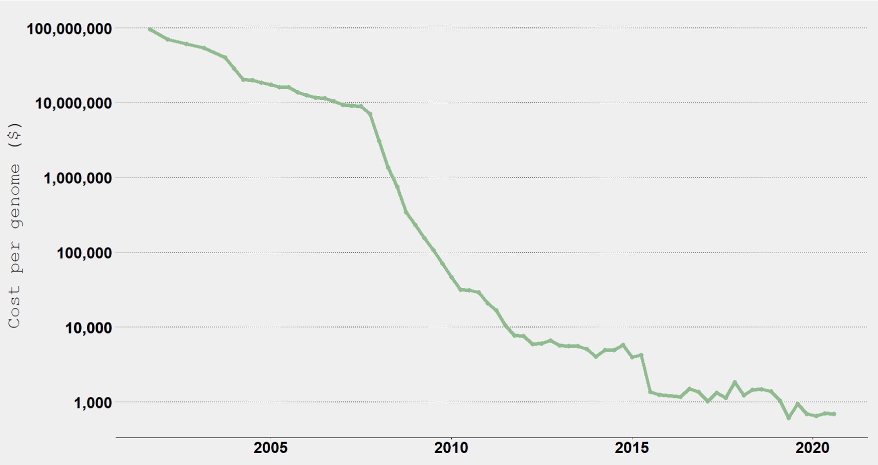


Figure 1. Cost per human genome, in US$, from the first draft until today. From 2001 to October 2007, Sanger sequencing cost to 6X coverage was used. From January 2008 to August 2020, Illumina sequencing cost to 30X coverage was used. Prepared from NIH data (https://www.genome.gov/sites/default/files/media/files/2020-12/Sequencing_Cost_Data_Table_Aug2020.xls).

Types of genetic variation

The term genetic variation –or, for simplicity, variation– refers to the differences in the DNA sequences between individuals or between populations. In this context, it is important to understand that, although they are sometimes used as synonyms, mutation and variant are two different concepts. Mutation is the biological process by which a genetic variant originates. And a variant, as mentioned, is something that differs between two given DNA sequences. In genomics, the word variant is commonly used when referring to a difference between one or several individuals and the reference genome used. In the general population, the term mutation is normally viewed with negative connotations. This use started around 1956, with the genetic damage reported from nuclear radiation (Condit et al. 2002) and it is still present in our societies nowadays.

Human genetic variation has been examined for a long time using traditional methods, like the study of microsatellites (Jarne and Lagoda 1996), but the availability of a reference genome and the expansion of MPS techniques have facilitated and boosted this process.

Genetic variants can be classified according to their size. From smaller to larger, we find single nucleotide variants (SNVs), small insertions and deletions (indels) and structural variants (SVs). Within the last category there are copy number variants (CNVs), insertions, deletions, duplications, and inversions (Figure 2). Indels and SVs were differentiated by their size: historically, events with < 1 kb were called indels (Feuk, Carson, and Scherer 2006) but, more recently, the threshold has arbitrarily been placed at 50 bp (Tattini, D'Aurizio, and Magi 2015).

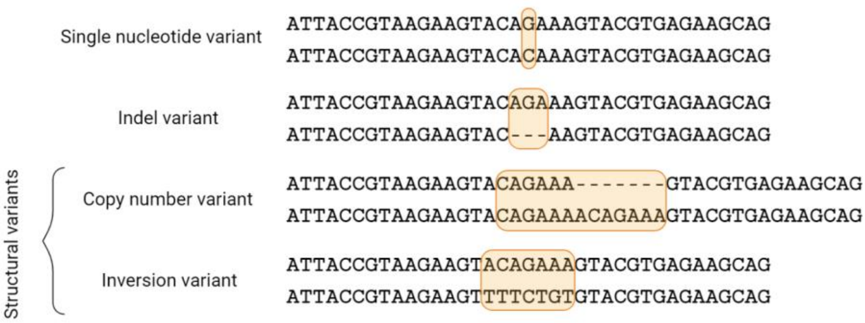

Single nucleotide variant
ATTACCGTAAGAAGTACAGAAAGTACGTGAGAAGCAG
ATTACCGTAAGAAGTACACAAAGTACGTGAGAAGCAG
Indel variant
ATTACCGTAAGAAGTACAGAAAGTACGTGAGAAGCAG
ATTACCGTAAGAAGTAC---AAGTACGTGAGAAGCAG
Structural variants
Copy number variant
ATTACCGTAAGAAGTACAGAAA-------GTACGTGAGAAGCAG
ATTACCGTAAGAAGTACAGAAAACAGAAAGTACGTGAGAAGCAG
Inversion variant
ATTACCGTAAGAAGTACAGAAAGTACGTGAGAAGCAG
ATTACCGTAAGAAGTTTTCTGTGTACGTGAGAAGCAG

Figure 2. Possible types of genetic variation in the human genome. Toy example using the coding sequence of exon 3 of *NLRP3* gene. Inspired by (Frazer et al. 2009). Created with BioRender.

SNVs have traditionally been the most studied type of variation, while SVs are harder to discover and their study has been more limited. In modern times, their role has increasingly been acknowledged: SVs encompass a longer proportion than SNVs in a given individual genome and they also span longer coding regions (Redon et al. 2006).

Germline and somatic variants

Depending on the moment when the mutation occurred, we can differentiate between germline and somatic variants. Somatic variants are caused by postzygotic mutations, those arising after the fertilization of the ovule by the spermatozoid. These mutations can occur from the beginning of the development of an organism to any time point during its lifespan. Postzygotic mutations give rise to mosaicism, in which different cells of the same individual present dissimilarities in their genomes (Lupski 2013). In contrast, germline variants are inherited by the individuals directly from their parents. There are two types of germline variants, homozygous, when both inherited copies are the same, and heterozygous, when they are different.

Germline SNVs are clearly the best known type of variation, and they have been thoroughly studied for decades. In contrast, somatic variants have started being investigated more recently, mostly in relation to cancer, since they are the main driving force of its appearance and development (Alexandrov et al. 2020). This has been deeply studied in the past years, with great efforts such as The Pan-Cancer Analysis of Whole Genomes (Campbell et al. 2020), and it has been extensively reviewed elsewhere (Alexandrov and Stratton 2014; Martínez-Jiménez et al. 2020; Martincorena and Campbell 2015). Considering somatic variants and non-cancer diseases, the first review on the topic was published in 2013 (Li and Williams 2013).

Somatic variants accumulate with time, as observed in multiple tissues and cell types, for instance, in the epithelium of the oesophagus (Martincorena et al. 2018; Yokoyama et al. 2019), endometrium (Moore et al. 2020), lung (Yoshida et al. 2020), colon (Nicholson et al. 2018; Lee-Six et al. 2019) and bladder (Lawson et al. 2020). Also in non-epithelial cells such as bulk skin (Martincorena et al. 2015) and individual melanocytes (Tang et al. 2020), liver (Brunner et al. 2019), neurons (Lodato et al. 2018) and blood (Genovese et al.

2014; Jaiswal et al. 2014; Xie et al. 2014; Zink et al. 2017). Somatic variation happens in mature cells, but it has also been observed in adult stem cells of the small intestine, colon and liver (Blokzijl et al. 2016). In fact, the rate of accumulation of these mutations is accelerated when the aging process is boosted in mice (Odagiri et al. 1998).

Somatic variation has a particularly important role in blood. Clonal haematopoiesis (CH) happens when a single clone carrying a somatic variant expands exponentially and becomes overrepresented in blood. It is related to aging and several malignant pathologies. In a set of studies published in 2014 (Genovese et al. 2014; Jaiswal et al. 2014; Xie et al. 2014), several genes were linked to CH, although in an earlier work from 2012 the most important ones were already mentioned (Welch et al. 2012). We can highlight *DNMT3A*, *ASXL1* and *TET2*, three epigenetic regulators, as genes recurrently found to harbour somatic variants in individuals with CH. Recently, and taking advantage of a dataset generated for cancer research with around 12,000 blood-cancer paired samples, the list of genes under positive selection in CH has been expanded up to almost 70 (Pich et al. 2020).

In the past years, CH has been described as even more common than previously observed in people older than 65 (Zink et al. 2017). In that study, genetic drift was proposed as the process behind CH, something that also happened with clonality in other tissues like the stomach (Barker et al. 2010) and intestine (Lopez-Garcia et al. 2010). However, in a metastudy published in 2020, in which several of the aforementioned big datasets were reanalysed, positive selection was determined to be the main driving force of CH (Watson et al. 2020). The relationship between CH and disease is also explored in the previous works. Not surprisingly, different haematological cancers are the most linked diseases to this process. All of this has been reviewed elsewhere (Jaiswal and Ebert 2019).

Mutations as source of evolution, diversity… and disease

Mutation is the ultimate source of genetic variation and diversity and thus, of evolution (King, Soller, and Kashi 1997). But, on the other side of this coin, when deleterious mutations happen, we can encounter diseases or even lethal mutations that are not compatible with life.

To briefly comment on genetic diversity in humans, it is essential to mention the relevant efforts that have shed some light on this topic. There are studies including individuals from populations all over the globe, such as the HapMap Project (International HapMap Consortium 2003), the 1,000 Genomes Project (The 1000 Genomes Project Consortium 2015) or the Simons Genome Diversity Project (Mallick et al. 2016). There are also local studies such as the 100,000 Genomes Project from Genomics England or the Icelandic project (Gudbjartsson et al. 2015). A review of the main projects until 2017 was published (Hindorff et al. 2018). Although extremely interesting, evolution and diversity are out of the scope of this thesis, which is focused on disease.

In this regard, genetic diseases are caused by an abnormal expression of one or several genes. Monogenic diseases are those in which only one mutated gene is responsible for the pathogenic phenotype. Examples of this model are Huntington disease (*HTT*), haemophilia A (*F8*) or Duchenne muscular dystrophy (*DMD*). Due to their heritability, they are also called Mendelian diseases (Gilissen et al. 2011). They can be classified according to their type of inheritance. This way, we find autosomal dominant (AD), autosomal recessive (AR), X-linked dominant, X-linked recessive and Y-linked diseases.

However, the majority of diseases are caused by genetic variants affecting not one but multiple genes. These are known as polygenic disorders. Besides, the environment is a very important factor in most of these cases, so they are also called complex diseases (Badano and Katsanis 2002). Famous examples are Alzheimer's disease, obesity or type 2 diabetes. Other scenarios, like the digenic (Gazzo et al. 2017; de Valles-Ibáñez et al. 2018) or the oligogenic model, which contemplates few genetic modifiers (Kousi and Katsanis 2015), are being explored in recent times.

Autoinflammatory diseases, which are the main focus of this thesis and that will be further detailed, present several of the models above mentioned depending on the clinical entity. There are monogenic autoinflammatory diseases with autosomal dominant inheritance –e.g. Muckle-Wells syndrome (Muckle and Wells 1962)–, autosomal recessive inheritance –e.g. deficiency of IL-1 receptor antagonist (DIRA) (Aksentijevich et al. 2009)–, X-linked dominant –e.g. X-linked reticulate pigmentary disorder (Starokadomskyy et al. 2016)–, digenic –e.g. proteasome-associated autoinflammatory syndrome (Brehm et al. 2015)– and polygenic –e.g. inflammatory bowel disease (Loddo and Romano 2015)–.

1.2. How to detect genetic variation: MPS data

Introduction to MPS data

MPS has been the most used strategy to generate data for the study of genetic variation during the last 15 years. Short and high quality Illumina reads have demonstrated their value in deciphering SNVs, indels and SVs, especially deletions, in which the absence of coverage at a certain locus points to such events. Long read strategies, mainly Nanopore and PacBio, offer less per-base accuracy, but they have improved both the generation of new assemblies and the calling of SVs. Their value has already been proven in the field of medical genomics (Mantere, Kersten, and Hoischen 2019).

The first step in any MPS experiment is the extraction of DNA from the sample. After that, different types of DNA library preparations are needed depending on the technology used. This step is used to arrange the genetic material in a particular way to prepare it for sequencing.

When the Illumina sequencing technology is selected, first, the DNA is fragmented to a specific size and special adapters are ligated at its ends in the library preparation step (Meyer and Kircher 2010). PCR amplification of the material is usually needed at this point. Then, cluster amplification can start. In this process, the library is loaded into a flow cell and each fragment is amplified into a clonal cluster. Finally, the actual sequencing, which is called sequencing by synthesis, takes place. The sequencer detects the DNA bases when they are incorporated into the template strands and they are distinguished by the use of fluorescent labelled deoxyribonucleotide triphosphates (dNTPs) (Bentley et al. 2008). Single-read or paired-end sequencing strategies can be chosen. In the case of paired-end, both ends of a fragment are sequenced, which can be useful to detect genomic rearrangements. Then, the base calling process transforms the signals of the machine into raw reads in FASTQ format.

We can find three types of generated data depending on the proportion of the genome sequenced: WGS, WES and targeted sequencing (Figure 3). The decision of choosing one over another will depend on the aim of the study, but also on the available budget, computational resources and specialized personnel. In that regard, WGS is more expensive and the generated files are usually bigger than those for WES and targeted sequencing.

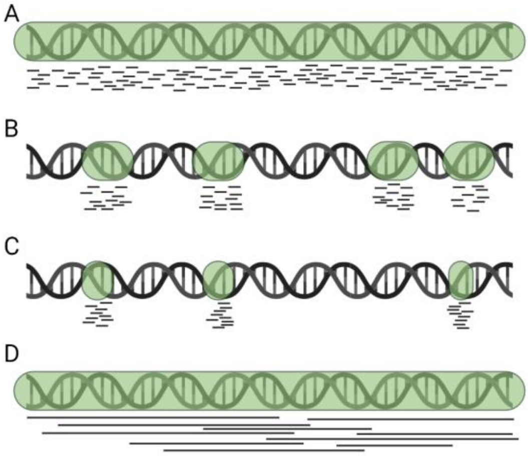


Figure 3. Types of MPS data. **A.** Short reads in WGS. **B.** Short reads in WES. **C.** Shor reads in targeted sequencing. **D.** Long reads in WGS. Created with BioRender.

WGS reaches the highest resolution by sequencing all nucleotides of the sample. This technique is used to discover SNVs, indels or SVs all over the genome or in non-coding regions. WES is intended to search for coding variants and, although specific software has been developed to study SVs (Fromer et al. 2012; Krumm et al. 2012), it is not the recommended strategy due to the high fragmentation of the data. Besides, it has been observed that WGS is more powerful than WES when calling variants in exonic regions (Belkadi et al. 2015; Meienberg et al. 2016). However, WES is more cost-effective than WGS and higher coverages can be reached more easily, which is particularly interesting in somatic variant discovery. Targeted sequencing is the most cost-effective strategy, since only predefined regions of the genome are sequenced.

PacBio and Nanopore data are generated differently. These third generation technologies are also called single-molecule sequencing (SMS), because PCR amplification of the genetic material is not required and, in this way, single whole DNA molecules can be sequenced.

The PacBio technology was the first third generation technology to be developed, and its bases were defined already in 2009 (Eid et al. 2009). This technology adapts the sequencing by synthesis approach used by Illumina and it is based on wells where the DNA is attached. Each molecule is labelled with a fluorescent dye and their signals are detected (Rhoads and Au 2015). ONT released its first instrument, the MinION, in 2014. Their technology takes advantage of changes in the electrical conductivity generated by DNA molecules when passing through biological nanopores (Lu, Giordano, and Ning 2016). These changes are registered and they can afterwards be translated into DNA bases. Besides, it is also aware of modifications in the DNA like methylation, so it can be used in epigenomic analyses (Simpson et al. 2017). PacBio technology can also detect these modifications (Davis, Chao, and Waldor 2013; Feng et al. 2013), but it is not commonly done due to its higher cost compared to ONT (Gouil and Keniry 2019).

Calling SNVs and indels using Illumina data

SNV and indel calling is the process by which this type of variation is derived from a MPS experiment. Once the raw Illumina data is generated, it is common to use the GATK (Genome Analysis Toolkit) best practices workflow (McKenna et al. 2010) for this purpose (Figure 4). Briefly, in this pipeline, raw reads in FASTQ format are mapped to a reference genome using BWA-MEM (Li 2013) algorithm. This generates the aligned SAM/BAM files, which are then explored to mark duplicates (artefact reads originated by the techniques used in generating the data) and to perform base quality score recalibration (BQSR). BQSR detects systematic errors in the sequencing process and it recalculates the qualities of the base calls. Then, GATK provides its own variant caller for germline SNVs and indels: HaplotypeCaller. The variant calling process generates VCF (variant calling format) files, which have to be filtered to eliminate erroneous calls based on their quality, the read depth or other parameters such as the mappability of the data.

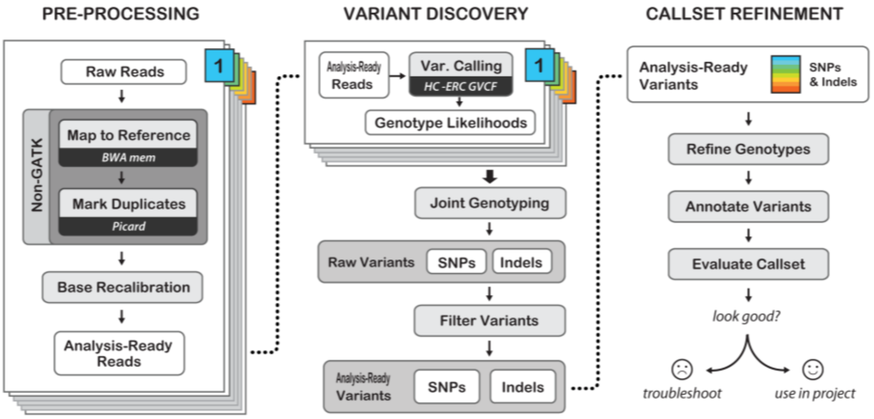


Figure 4. GAKT's best practices workflow for germline SNVs and indels (https://software.broadinstitute.org/gatk/best-practices/).

The special scenario of somatic variation

The case of somatic variants is particular due to the lower variant allele frequency (VAF) they present in the tissue. As previously mentioned, germline variants can be homozygous or heterozygous. In a sequencing experiment, the VAF of a germline heterozygous variant would be ~50%, since half of the reads would support the alternative allele, although some deviations from this expected value may occur. In contrast, somatic variants normally present VAFs < 50% –as low as 1% and below–. In this scenario, different strategies are used to call them, since algorithms designed for germline variants would fail in this task.

Besides, somatic variants have been widely studied in cancer genomics, in which a tumour sample is compared with a healthy tissue, commonly blood, that serves as a background for removing the basal germline variation (Cai et al. 2016; Hofmann et al. 2017; Krøigård et al. 2016; Xu et al. 2014). In the cases in which this comparison cannot be established, strategies using just a tumour sample have been developed (Sandmann et al. 2017; Teer et al. 2017). But there is an increasing interest in studying somatic variation in healthy tissue (Wang et al. 2021) and in diseases other than cancer (Van Horebeek, Dubois, and Goris 2019), and thus, new laboratory techniques and software are being developed. For example, the use of unique molecular identifiers has emerged, for which smCounter2 has been particularly made to call variants (Xu et al. 2019). It is also possible to call somatic variants taking advantage of certain characteristics of the data such as the allelic imbalance (Luquette et al. 2019).

MuTect2 is the software to call somatic variants provided by GATK. Apart from it, there are several different programs intended for this use: CaVEMan (Jones et al. 2016), EBCall (Shiraishi et al. 2013), LoFreq (Wilm et al. 2012), SomVarIUS (Smith et al. 2016), Strelka2 (Kim et al. 2018), VarDict (Lai et al. 2016), VarScan2 (Koboldt et al. 2012), etc. However, the comparison of the results generated by these tools shows poor levels of overlap (Cai et al. 2016; Krøigård et al. 2016). To solve this issue, pipelines combining some of the aforementioned strategies have been proposed (Callari et al. 2017; Kim, Jacob, and Speed 2014).

After somatic variant calling, and depending on the generated data, the application of a set of filters to reduce the list of candidate variants is necessary. Because of the low VAFs, the number of candidate genetic variants is usually very high and true somatic variants need to be discriminated from germline variants and sequencing or mapping errors. The commonly used filters are based on the quality of the mapping and calling, the expected VAFs, the comparison of the sample with other individuals or other samples from the same one but from different tissues (Lobon et al. 2020), etc.

Variant annotation

Once the VCF files containing the candidate germline, somatic variants or both have been generated, several tools can be employed to annotate different characteristics. This is a common practice in clinical genomics, when the aim is to discover pathogenic variants. We highlight Ensembl's VEP (variant effect predictor) (McLaren et al. 2016) and SnpEff (Cingolani et al. 2012). These tools classify the variants according to their predicted impact in the protein in high, moderate, low and modifier. High impact variants are those predicted as disruptive, such as the stop gained or frameshift variants. Loss of function variants, which produce a defective gene product, are classified within this first category. Moderate impact variants are mainly missense variants (they change the amino acid in the resultant protein), while low impact variants are the synonymous ones (they do not change the amino acid). Modifier impact variants are those located in non-coding regions of the genome and, as such, it is more difficult to estimate their potential effect.

Apart from this basic information, there are other parameters that help to provide a better understanding of the variants. For instance, bioinformatics predictors such as SIFT (Vaser et al. 2016) or PolyPhen-2 (Adzhubei et al. 2010) are commonly used in describing the potential pathogenicity of SNVs. These tools give a score and classify the variants according to their predicted effect in the protein. SIFT has two categories that are deleterious and tolerated, while PolyPhen-2 has four, probably damaging, possible damaging, benign and unknown. Another interesting parameter is the CADD (combined annotation dependent depletion) score (Rentzsch et al. 2019), which is a compendium of more than 60 other metrics. CADD simultaneously accounts for conservation parameters, epigenetic modifications, functional predictions and the genetic context of the variants. Then, it ranks all possible SNVs in the human genome and it creates a C-score for them. Although arbitrary, a threshold value of 15 is normally considered to identify potential pathogenic variants. A combination of several of these tools is usually used when performing pathogenic variant discovery.

Another important aspect to take into account when dealing with clinical genomics data is the minor allele frequency (MAF). MAF is defined as the allele frequency of a variant in a determined sequencing experiment or population. If a rare monogenic disease is being studied, the MAF of the potential candidate variant in the general human population needs to be consistently rare. A powerful resource in this sense is the Genome Aggregation Database (gnomAD). The current version of the project, v3.1, harbours allele frequency data from 76,156 WGS, while in the previous version, v2.1, there were already 125,748 WES and 15,708 WGS (Karczewski et al. 2020).

Consider, as a practical example, that we are dealing with a patient with cryopyrin associated periodic syndromes, which are monogenic autoinflammatory disorders with an estimated prevalence of 1-2 cases in 1,000,000 people in the USA and Europe (Martorana et al. 2017). If we find an interesting candidate variant, with a SIFT prediction of deleterious, a PolyPhen-2 score of probably damaging and CADD > 20, but we observe in gnomAD that its MAF is 20% in those populations, it is not possible that it is the causal one.

After the variant calling and the filtering processes, manual and visual inspection of the mapped reads is a necessary step to ensure the robustness of the candidate variants. This approach gives a general idea of the variant in its genomic context and it can serve as an additional filter. The Integrative Genomics Viewer (IGV) (Robinson et al. 2011) has proven to be a valuable tool for this purpose. Besides, validation through orthogonal approaches such as Sanger sequencing is necessary. In the somatic scenario, Sanger sequencing is normally not enough, since this technique is blind to VAFs < 20%. In these cases, amplicon-based deep sequencing is being used in current studies, also providing a more accurate estimation of the VAFs due to the high coverages it achieves (> 20,000X). Of note, the simultaneous analysis of different amplicons encompassing the same genetic variant is recommended to avoid inaccurate frequency estimations because of PCR biases (Mensa-Vilaró et al. 2018).

Calling structural variants using Illumina data

There are several strategies to call SVs in MPS data. Among them, we find the split-reads based methods, paired-end mapping methods, read depth methods or a combination of the last two of them (Zhao et al. 2013). Read depth methods are the most popular ones. They take advantage of the coverage of the samples and use statistical models to call structural events.

WES and targeted sequencing are not the most suitable solution to explore this type of variation due to the high fragmentation of the data, but specific software has been developed for that purpose. CoNIFER (Krumm et al. 2012), for instance, requires the use of an aligner like mrsFAST (Hach et al. 2014) that maps the reads to multiple locations in the genome, and then it calculates RPKM (reads per kilobase million) and corrects these values with a Z transformation and a singular value decomposition transformation. XHMM (Fromer et al. 2012) uses principal component analysis to normalize the read depths of the samples and then a hidden Markov model to perform the SV calling.

For WGS data, other different tools have been developed (Zhang et al. 2019). The most cited are mrCaNaVAR (Alkan et al. 2009) and CNVnator (Abyzov et al. 2011), which use a read depth strategy.

Similar to the SNVs and indels, visual inspection of the mapped reads can help to elucidate if the detected structural variant is actually an artefact caused by mapping or other errors. Again, validation through orthogonal approaches, in this case, for instance, qPCR (quantitative PCR) or aCGH (array comparative genomic hybridization) (Alkan, Coe, and Eichler 2011), is necessary.

Calling structural variants using Nanopore data

Although Nanopore and PacBio's long reads can also be used for SNV calling, their error rate is still as high as 15%, with ONT errors being more complex than the PacBio ones (Rang, Kloosterman, and de Ridder 2018). New methods with promising results are being thoroughly investigated in this sense, for example, leveraging the haplotype information of these reads (Edge and Bansal 2019) or combining short Illumina reads with long reads (Holley et al. 2021). So far, the main uses of long reads has been the *de novo* generation of assemblies –especially interesting for non-model organisms– and structural variant calling (Pollard et al. 2018).

Similar to the short reads scenario, FASTQ files containing the raw reads are generated after the base calling of the raw data obtained from the sequencing machines. Then, although BWA is also capable of mapping long reads, specific software have been developed for this task, such as Minimap2 (Li 2018) or NGMLR (Sedlazeck et al. 2018).

After mapping, several programs have been released for the study of SVs in long read data. To mention just a few, we can highlight Sniffles (Sedlazeck et al. 2018) and SVIM (Heller and Vingron 2019). Sniffles uses a technique that combines split-read alignments, high mismatch regions and coverage analysis to call SVs. SVIM uses a three-step method: collection of SV signatures, cluster by genomic position and span and combination and classification of the SV signature clusters.

It is also possible to combine both of the main uses of long reads, since we can generate an assembly and call SVs in it in comparison to another assembly. For this purpose, it is first necessary to generate the assembly with tools like

Canu (Koren et al. 2017) and then to align it to another chosen assembly by using, for instance, nucmer tool in MUMmer (Eisen et al. 2000). For the structural variant calling process, other tools like Assemblytics (Nattestad and Schatz 2016), that calls the SVs based on their distinct alignment signatures, and SVIM-asm (Heller and Vingron 2020) have been developed.

Powerful SVs callsets are being generated and compared to those based on short reads (Audano et al. 2019), establishing an improvement in the resolution for this type of variation (Mahmoud et al. 2019). The visual inspection of the mapped reads and the validation through orthogonal approaches is also applicable with this type of data.

Flow sorting enrichment of individual chromosomes

Flow cytometry is a technique that detects and measures different characteristic of the samples –mainly cells, but also other particles–, such as their dimensions, their complexity, the number of events, their speed when passing through a capillary tube and others (Picot et al. 2012). The most rudimentary flow cytometer was described in 1934 (Moldavan 1934) and, since then, the technology has been greatly refined.

One of the applications of flow cytometry is flow sorting. Flow sorting is the process by which different populations from a sample can be identified according to certain parameters and, thus, they can be physically separated. Interestingly, this methodology can also be used to separate individual chromosomes (Gray et al. 1975). More recently, chromosome sorting has been improved to the point of obtaining millions of individual chromosomes from a cell line. These chromosomes can be then sequenced to study their structural variant landscape (Kuderna et al. 2019). This way, the exploration of SVs in a particular chromosome is facilitated, since higher coverages can be reached with no need to sequence the whole genome of the samples.

2. Immune system and autoinflammatory diseases

2.1. The immune system

Historical overview

The immune system is the compendium of complex biological processes that vertebrates possess to defend themselves from disease, regardless of the aetiology. These types of mechanisms act against external agents and also against some internal threats (e.g. cancer).

Our understanding of immunological processes was first pioneered by the scientific discoveries of Elie Metchnikoff and Paul Ehrlich (Figure 5), who both received the Nobel Prize in Physiology or Medicine in 1908 for their work. On the one hand, Metchnikoff discovered phagocytes and the phagocytosis process and, on the other hand, Ehrlich postulated the side-chain theory (the lock-and-key principle) that works for immunoglobulins (antibodies) and T cell receptors (TCRs). By doing this, they set the basis for what would become the division of the immune system in innate and adaptive (Kaufmann 2008).

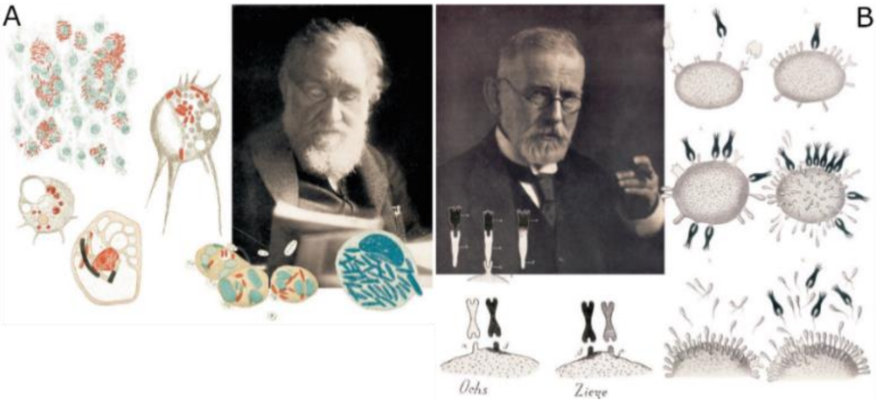


Figure 5. The founders of immunology. **A.** Elie Metchnikoff and his drawings of the phagocytic process of bacteria by micro- and macrophages. **B.** Paul Ehrlich and his drawings of antibodies. Adapted from (Kaufmann 2008).

The innate and the adaptive immune system are conceptually differentiated by their function, response time, cell type composition and evolutionary origin.

We can trace back the appearance of innate immune systems around 800 million years ago, with the origin of Metazoan, i.e., animals (Figure 6), although this date is not clear (Cunningham et al. 2017). Therefore, all animals present innate immunity with its key features as phagocytes, the complement system and pattern recognition receptors (PRRs).

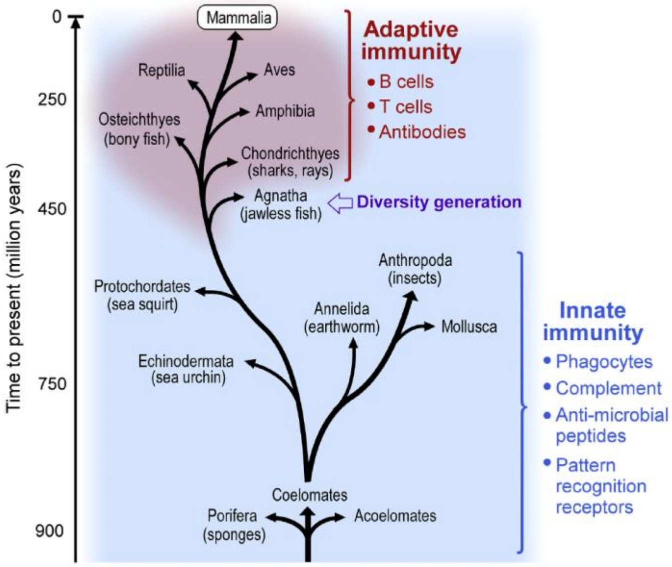


Figure 6. The immune system in the tree of life of Metazoan (Nigrovic, Lee, and Hoffman 2020).

The adaptive immunity, characterized by the somatic recombination of immunoglobulin and TCR genes, was thought to have first evolved in the ancestor of jawed fishes (Matsunaga and Rahman 1998; Bartl et al. 2003). Unexpectedly, a similar process has been described in the sea lamprey, a jawless fish, but in this case by taking advantage of rearrangements of leucine-rich repeats modules in the *VLR* (variable lymphocyte receptor) locus (Pamcer et al. 2004), placing the origin date around 450 million years ago. If the reader is interested, reviews on the evolutionary history of adaptive immunity can be found in the following articles (Cooper and Alder 2006; Flajnik and Kasahara 2010).

Innate and adaptive immune system

In general terms, the innate immunity is characterized by having an instantaneous and non-specific response to pathogens through general receptors. In contrast, the adaptive immunity is slower and it is in charge of developing a specific response and the immunological memory against the pathogen. This way, it prepares the organism to react faster and more efficiently in case of reinfections.

In mammals, the process by which the immune cells (leukocytes), along with the rest of blood cells, are formed is called haematopoiesis (Figure 7). In adults, it mainly takes place in the bone marrow and it begins in the pluripotent haematopoietic stem cell (HSC). From it, two progenitors are derived: the common myeloid progenitor (CMP) and the common lymphoid progenitor (CLP).

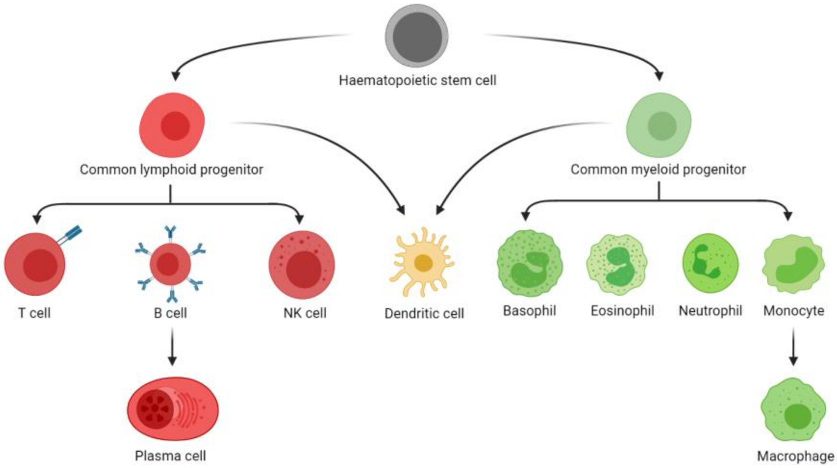


Figure 7. Simplified haematopoiesis process showing the main types of leukocytes. Created with BioRender.

Generally, cells derived from the CMP mediate the innate immunity, whereas those arising from the CLP are involved in the adaptive immunity. In the first group, we find granulocytes (neutrophils, eosinophils and basophils), mast cells, monocytes and macrophages and dendritic cells. In the second group, we find B and plasma cells and T cells. Natural killer cells show behaviours shared by both innate and adaptive immune responses, although they are mainly classified within the innate.

Somatic recombination, a concept that was above introduced, is the process by which mutations occur in somatic cells and these genetic alterations are transmitted to the daughter cells. It is opposed to the event that happens during meiosis and the formation of gametes. In immunology, somatic recombination is the physiological process that assembles immunoglobulin and TCR genes in the development of the lymphoid lineage and gives rise to their considerable diversity (Gellert 1992).

Genes encoding the receptors of the innate immune system, the PRRs, do not suffer somatic recombination, so their main strategy is to recognize conserved patterns: pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs). PAMPs are present across large groups of microorganisms and a well-known example are the bacterial lipopolysaccharides (LPS) (Janeway 1989), while DAMPs are present in the same tissues of the host.

PAMPs are recognized by the PRRs of macrophages, dendritic and B cells, the antigen-presenting cells. Then, mainly macrophages phagocyte the pathogens, which are introduced in the lysosomes to be degraded to peptides. These peptides are presented through the major-histocompatibility complex (MHC) molecules on their surface, which attract the attention of lymphocytes and, upon their activation, the adaptive response begins (Figure 8). At the same time, signalling receptors recognize the aforementioned PAMPs and induce the expression of several genes, including type I interferons (IFN-α and IFN-β), proinflammatory cytokines like IL-1 and other cytokines that recruit and activate cells of the adaptive immunity (Medzhitov and Janeway 2000).

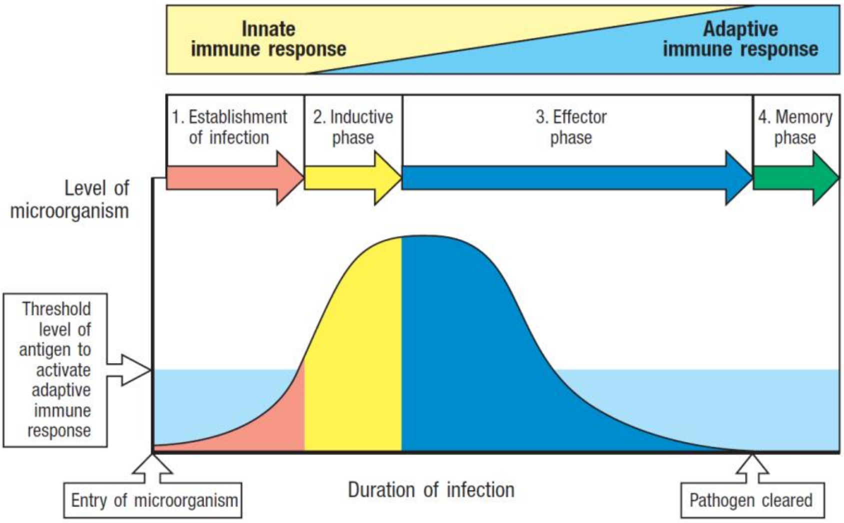


Figure 8. Diagram showing the sequential nature of the immune response in addressing infection (Murphy and Weaver 2017).

Lymphocytes are in charge of adaptation and memory and, because of them, the immune response is increasingly specific and reinfections are cleared more efficiently. Immunoglobulin and TCR genes do undergo somatic recombination and, if they are proven to be efficient against a certain antigen, their producer cells are clonally selected and expanded (Burnet 1976).

B cells produce antibodies that are attached to their membranes and recognize antigens. Once they are selected, they are transformed into plasma cells that produce and secrete that particular antibody in large amounts. These antibodies bind to the pathogen and its toxic products. T cells show TCRs in their membranes and follow a similar selection process. T cells can be subdivided in CD4$^+$ (T$_H$ or helpers and T$_{reg}$ or regulators) and CD8$^+$ (cytotoxic).

The division of the immune system in innate and adaptive, although practical for the study of immunology and for this thesis, is usually not that clear, as both systems work together and are supported by each other. Actually, the observation that certain cell types of the innate immune system have memory properties already created the debate of whether or not it is still rigorous to do this separation (Lanier and Sun 2009). The immune system is thoroughly regulated, since its homeostasis is vital for the organism, the excess or the defect of any of its parts being cause for disease.

Inflammation

Inflammation is a physiological process, mounted by the innate immune system, that acts in response to infection, tissue damage and tissue stress. Its role is to protect the organism against pathogens and induce the repair of the affected tissue to achieve the homeostasis. It starts with an acute phase characterized by the extravasation of leukocytes, mainly neutrophils, to the affected area, where they are activated. Then, they release reactive oxygen and nitrogen species, proteinase 3, cathepsin G and elastase to eliminate the disturbance. Tissue damage also occurs at this stage due to the toxicity of these components. If the inflammation persists, neutrophils are replaced by macrophages and, in the case of dealing with an infection, also T cells are recruited (Medzhitov 2008).

In a normal situation, inflammation is resolved once its function is fulfilled. Of note, the mechanisms of inflammation induced by infections are better known than others and, besides, systemic inflammation does not follow the same pattern from acute to chronic local inflammation (Medzhitov 2008). This information is relevant to understand autoinflammatory diseases, the main focus of this thesis.

Inflammasomes are a set of proteins assembled in the cytosol that play a major role in the inflammation process. They are called after the main protein forming them and, among the canonical inflammasomes, we find the pyrin, cryopyrin (NLRP3), NLRC4, NLRP1 and AIM2 inflammasomes. Canonical inflammasomes recruit the inactive pro-caspase-1, which is cleaved into active caspase-1 (Yang, Chang, and Baltimore 1998). At the same time, active caspase-1 is a protease that cleaves the precursor cytokines pro-IL-1β and pro-IL-18 into their active forms, IL-1β and IL-18, apart from producing pyroptosis, a type of inflammatory cell death (Guo, Callaway, and Ting 2015).

IL-1β and IL-18 are members of the IL-1 family of cytokines. There are 11 members in this family: IL-1α, IL-1β, IL-1Ra, IL-18, IL-33 and IL-1F5–IL-1F10 (Sims and Smith 2010). IL-1 was the first described cytokine (Gery, Gershon, and Waksman 1972) and, at that time, it was called lymphocyte-activating factor.

IL-1β is known to induce fever (Horai et al. 1998) and promote local effects like the activation of vascular endothelium and lymphocytes, local tissue destruction and increase of effector cells to the area of interest (Murphy and Weaver 2017). IL-18 induces the expression of INF-γ (Nakamura et al. 1989) and stimulates naïve T cells and NK cells (Yasuda, Nakanishi, and Tsutsui 2019). Only IL-1Ra (IL-1 receptor antagonist), encoded by *IL1RN* and described in 1987 (Seckinger et al. 1987), is a classical signal peptide. It binds to the IL-1 receptors to antagonize competitively the inflammatory effects of IL-1α and IL-1β (Dayer, Oliviero, and Punzi 2017).

The NLRP3 as a model of inflammasomes

The most characterized inflammasome is NLRP3. The *NLRP3* gene is located in chromosome 1 (chr1:247,416,162-247,448,822 in GRCh38 coordinates) and it encodes the cryopyrin protein. This gene, described in 2001 (Hoffman et al. 2001), is expressed in all cell types of the immune system, especially in those responsible for the innate immunity. Macrophages show the highest expression levels according to The Human Protein Atlas data (Uhlén et al. 2015).

The NLRP3 inflammasome, like others, needs to be primed prior to its activation. This means that the expression of *NLRP3* is upregulated through NF-κB pathways, and also some post-transcriptional modifications are triggered to facilitate the regulation of the inflammasome complex. A classic example of priming signal would be the recognition of LPS by TLR4. However, recent discoveries show that priming is not necessary in monocytes *in vitro* (Gritsenko et al. 2020).

After the priming step, the NLRP3 inflammasome can be activated by a wide range of stimuli, both exogenous and endogenous. Exogenous stimuli include some crystals –alum, silica, asbestos– and bacterial toxins like nigericin. Endogenous signals comprise potassium efflux out of the cell, generation of mitochondrial reactive oxygen species, liberation of mitochondrial DNA and translocation of NLRP3 to the mitochondria among others (Sutterwala, Haasken, and Cassel 2014).

Once primed and activated, the NLRP3 inflammasome is able to start its activity. A schematic representation of these processes is depicted (Figure 9).

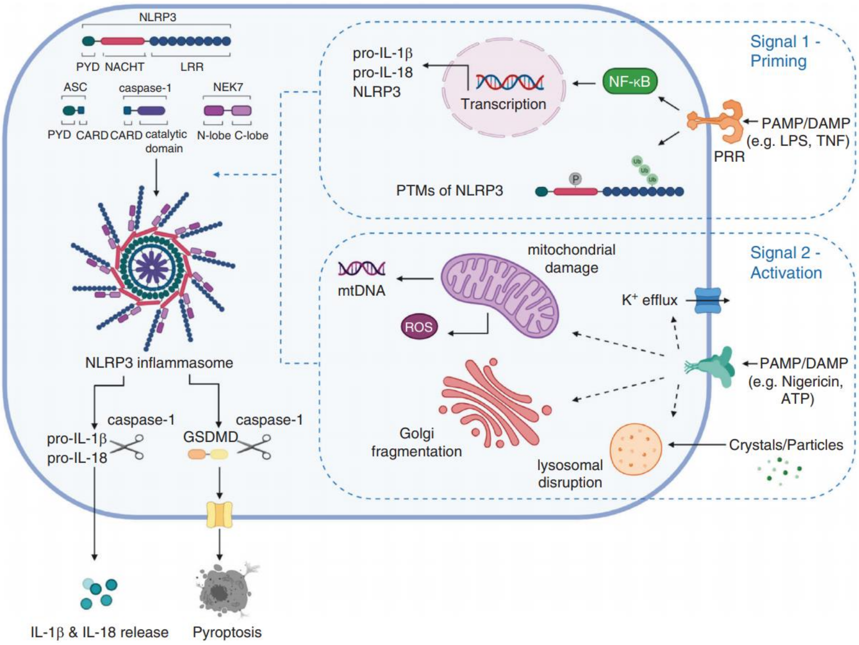


Figure 9. The NLRP3 inflammasome. Possible priming and activation signals, as well as its proinflammatory activity through the activation of caspase-1 (McKee and Coll 2020).

2.2. Autoinflammatory diseases

Overview of autoinflammatory diseases

Primary immunodeficiencies (PIDs) are a group of heterogeneous diseases caused by dysregulations of the immune system (Tangye et al. 2020). Autoinflammatory diseases are classified within PIDs. They can be briefly defined as those "in which the innate immunity plays the primary pathophysiologic role" (Manthiram et al. 2017). The problem with this simplified vision is that, as previously mentioned, innate and adaptive immunity are interconnected. While we could theoretically establish a link between innate–autoinflammation and adaptive–autoimmunity (and allergies), the reality is more complex. To further extend and detail the definition of autoinflammatory diseases, Nigrovic and Hoffman stated that the "pathogenic inflammation arises primarily through antigen-independent hyperactivation of immune pathways" (Nigrovic, Lee, and Hoffman 2020).

The first gene linked to an autoinflammatory disease was *MEFV*. This gene was described in 1996 (Touitou et al. 1996) and its relation to familial Mediterranean fever (FMF) was established one year after that (Santer et al. 1997). However, it was not until 1999 when the concept itself of autoinflammatory diseases appeared (McDermott et al. 1999).

The International Union of Immunological Societies (IUIS) is an organization that groups national and regional immunological societies from all over the globe. The IUIS publishes a biannual report updating the knowledge of PIDs, including autoinflammatory syndromes. In their latest version, a total of 45 different disorders and 42 genes known to harbour pathogenic variants have been listed for autoinflammatory diseases (Tangye et al. 2020), although in reality those are just the monogenic entities (Tables 1-4). The classification and number of disorders provided by the most recent review of Nigrovic and

Hoffman would be use in this thesis as reference (Nigrovic, Lee, and Hoffman 2020).

A blurry division line between autoinflammation and autoimmunity is observed in some cases. On the one hand, there are autoinflammatory disorders, such as some interferonopathies like the Aicardi-Goutières syndrome, that exhibit autoantibodies (Cuadrado et al. 2015). On the other hand, inflammation is common in autoimmune disorders, in which a genetic defect in adaptive immunity ends up activating inflammatory pathways. Although there are pure autoinflammatory and autoimmune disorders, reality also shows a continuum spectrum between both ends (Doria et al. 2012).

There are symptoms associated with autoinflammatory diseases that are shared by several of them. As a general consideration, systemic inflammation with recurrent fever episodes are common, as well as an elevation in acute phase reactants and a range of other manifestations such as rash, serositis and lymphadenopathy. Some clinical entities intercalate periods with symptoms with others with no symptoms. Others are chronic and skin manifestations such as dermatitis are common in them. The reasons behind the periodicity of some autoinflammatory diseases remain unclear (Georgin-Lavialle et al. 2020).

Groups of autoinflammatory diseases

Monogenic autoinflammatory diseases can be divided into four main different subgroups: inflammasomopathies, interferonopathies, disorders of the NF-κB and/or aberrant TNF activity and others (Nigrovic, Lee, and Hoffman 2020). Among the polygenic entities, we can highlight the Behçet disease, the inflammatory bowel disease and the systemic-onset juvenile idiopathic arthritis.

Monogenic entities follow a pattern of inheritance which can be either autosomal dominant or autosomal recessive. Additionally, defects in the

POLA1 gene cause the X-linked reticulate pigmentary disorder (Starokadomskyy et al. 2016). Autosomal dominant inheritance through gain of function mutations is relatively common, and two important examples are the *NLRP3* and *NOD2* variants causing the spectrum of cryopyrin associated periodic syndromes (CAPS) and Blau syndrome respectively (Manthiram et al. 2017).

Inflammasomopathies

Inflammasomopathies (Table 1) are caused by mutations in the genes forming the inflammasomes (*NLRP3*, *NLRC4*, *NLRP1*, *AIM2*, *MEFV*, etc.) or in genes encoding inflammasome regulators (*IL1RN* among others).

Table 1. Inflammasomopathies and other IL-1 family conditions.

Disease	Gene mutated	Inheritance	Somatic reported
FMF	*MEFV*	AR or AD	YES
PAAND	*MEFV*	AD	-
MKD	*MVK*	AR	-
PAPA	*PSTPIP1*	AD	-
Hz/Hc	*PSTPIP1*	AD	-
PFIT	*WDR1*	AR	-
FCAS	*NLRP3*	AD	-
MWS	*NLRP3*	AD	YES
CINCA/NOMID	*NLRP3*	AD	-
Majeed's	*LPIN2*	AR	-
AIFEC	*NLRC4*	AD	-
FCAS/NOMID	*NLRC4*	AD	YES
FCAS	*NLRP12*	AD	-
NAIAD	*NLRP1*	AD	-
DIRA	*IL1RN*	AR	-
DITRA	*IL36RN*	AR	-

FMF: familial Mediterranean fever; PAAND: pyrin-associated autoinflammation with neutrophilic dermatosis; MKD: mevalonate kinase deficiency; PAPA: pyogenic arthritis, pyoderma gangrenosum and acne; Hz/Hc: hyperzincemia/hypercalprotectinemia; PFIT: periodic fever, immunodeficiency, and thrombocytopenia; FCAS: familial cold autoinflammatory syndrome; MWS: Muckle-Wells syndrome; NOMID: neonatal-onset multisystem inflammatory disease; AIFEC: autoinflammation with infantile enterocolitis; NAIAD: *NLRP1*-associated autoinflammation with arthritis and dyskeratosis; DIRA: deficiency of IL-1 receptor antagonist; DITRA: deficiency of interleukin-36 receptor antagonist. Modified from (Nigrovic, Lee, and Hoffman 2020).

These diseases are produced by a hyperactivation of the inflammatory pathways mediated by inflammasomes, which leads to an uncontrolled production of caspase-1, IL-1β and/or IL-18, depending on the aberrant mechanism (Figure 10). Not surprisingly, IL-1 inhibitors (anakinra, canakinumab, rilonacept) are the treatment of election for many of these diseases (Lachmann et al. 2009). These pathogenic mechanisms through defects of the inflammasomes pathways are the best studied among autoinflammatory diseases (Georgin-Lavialle et al. 2020).

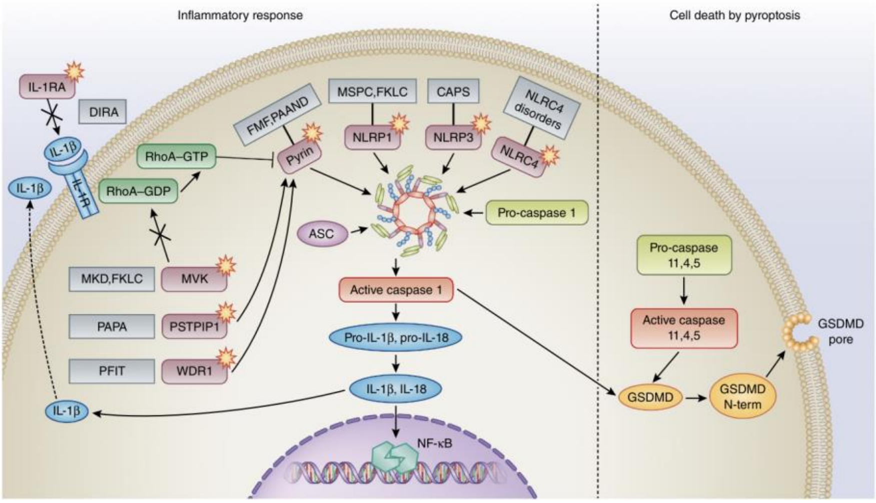


Figure 10. Mechanisms of action of inflammasomopathies and diseases caused by IL-1β. The affected proteins are in red boxes and the names of the diseases in grey boxes (Manthiram et al. 2017).

Some examples of inflammasomopathies are the aforementioned FMF, CAPS or DIRA (deficiency of IL-1 receptor antagonist).

FMF is caused by defects in the *MEFV* gene, which encodes the pyrin protein. CAPS are a set of diseases caused by defects in *NLRP3*. There are several different entities within CAPS: chronic infantile neurological cutaneous and articular (CINCA) syndrome –also called neonatal-onset multisystem inflammatory disease (NOMID)–, familial cold autoinflammatory syndrome (FCAS) and the Muckle-Wells syndrome (MWS). All of them present an autosomal dominant inheritance caused by gain of function mutations. These mutations hyperactive the pyrin and the NLRP3 inflammasomes in FMF and CAPS respectively.

DIRA, on the other hand, presents an autosomal recessive inheritance. The genetic defect is found in the *IL1RN* gene, which encodes the IL-1Ra (IL-1 receptor antagonist). One copy of this gene is enough to accomplish its function, which is to control the inflammation by antagonizing the effects of IL-1α and IL-1β (Dayer, Oliviero, and Punzi 2017), so the two copies must be mutated to cause the disease.

Interferonopathies

Interferonopathies (Table 2) are caused by a hyperactivation of the type I interferon (IFN-α and IFN-β) axis, which usually takes a role in antiviral defence. As with many other autoinflammatory diseases, patients with interferonopathies present fever, rash, systemic inflammation and skin vasculitis although, interestingly, many exhibit high titres of autoantibodies (Nigrovic, Lee, and Hoffman 2020).

Table 2. Type I interferonopathies.

Disease	Gene/s mutated	Inheritance	Somatic reported
Aicardi-Goutières syndrome	*TREX1, ADAR1, RNASEH2A/B/C, SAMHD1, IFIH1*	AR (AD: *IFIH1*)	-
Monogenic SLE	*DNASE1/2/1L3,* complements	AR (AD: *DNASE1*)	-
SMS	*IFIH1, DDX58a*	AD	-
SAVI	*TMEM173*	AD	YES
CANDLE / PRAAS, PRAID	*PSMB4, PSMA3, PSMB8, POMP, PSMG2,PSMB9, PSMB10*	Digenic, AR (AD: POMP)	-
AGS-like	*USP18, ISG15, STAT2*	AR	-
SPENCD	*ACP5*	AR	-

SLE: systemic lupus erythematosus; SMS: Singleton-Merten syndrome; SAVI: STING-associated vasculopathy of infancy; CANDLE: chronic atypical neutrophilic dermatosis with lipodystrophy and elevated temperature; PRAAS: proteasome-associated autoinflammatory syndrome; PRAID: POMP-related autoinflammation and immune dysregulation disease; AGS: Aicardi-Goutières syndrome; SPENCD: spondyloenchondrodysplasia. Modified from (Nigrovic, Lee, and Hoffman 2020).

Some examples of interferonopathies are the Aicardi-Goutières syndrome, caused by defects in several genes like *TREX1* or *ADAR1*, and SAVI (STING-associated vasculopathy of infancy), for which mutations in *TMEM173* are responsible.

Disorders of the NF-κB and/or aberrant TNF activity

Disorders of the NF-κB and/or aberrant TNF activity (Table 3) are mediated by the nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB). This protein complex is a key signalling hub that integrates multiple signals from inside and outside the cells. Among its many functions, it is a transcription factor that promotes the expression of proinflammatory genes. This way, the so-called NFκBopathies are characterized by fever, systemic inflammation and granuloma formation (Steiner et al. 2018). Important activators of NF-κB pathways are the TNF receptors, being the reason why they are grouped together in this category of autoinflammatory diseases.

Table 3. Disorders of the NF-κB and/or aberrant TNF activity.

Disease	Gene/s mutated	Inheritance	Somatic reported
HA20	*TNFAIP3*	AD	-
RELA haploinsufficiency	*RELA*	AD	-
ORAS	*OTULIN*	AR	-
LUBAC deficiency	*HOIL1, HOIP*	AR	-
Blau	*NOD2*	AD	YES
TRAPS	*TNFRSF1A*	AD	YES
DADA2	*ADA2*	AR	-
CRIA	*RIPK1*	AD	

HA20: haploinsufficiency of A20; ORAS: OTULIN-related autoinflammatory syndrome; HOIL1: heme-oxidized IRP2 ubiquitin ligase 1; HOIP: HOIL1-interacting protein; TRAPS: TNFR1-associated periodic syndrome; DADA2: deficiency of adenosine deaminase 2; CRIA: cleavage-resistant RIPK1-induced autoinflammatory syndrome. Modified from (Nigrovic, Lee, and Hoffman 2020).

In this group we find some diseases like the Blau syndrome, caused by defects in *NOD2*, or TRAPS (TNFR1-associated periodic syndrome), caused by mutations in *TNFRSF1A*.

Other autoinflammatory diseases

In this final set of other autoinflammatory diseases (Table 4) we find a myriad of clinical entities that present autoinflammatory processes. For example, mutations in the *COPA* gene originate malfunctions in the transport from the Golgi to the endoplasmic reticulum, which causes the COPA (COPI coat complex subunit alpha) syndrome, with symptoms like autoimmunity, inflammatory arthritis and lung damage (Watkin et al. 2015).

Table 4. Other mechanisms.

Disease	Gene/s mutated	Inheritance	Somatic reported
COPA	*COPA*	AD	-
PLAID	*PLCG2*	AD	-
APLAID	*PLCG2*	AD	-
SIFD	*TRNT1*	AR	-
LACC1 deficiency	*LACC1/FAMIN*	AR	-
VEO-IBD	*IL-10, IL10RA, IL10RB*	AR	-
ARPC1B deficiency	*ARPC1B*	AR	-
CDC42 deficiency	*CDC42*	AR	-

COPA: COPI coat complex subunit alpha; PLAID: PLCG2-associated antibody deficiency and immune dysregulation; APLAID: autoinflammation and PLAID; PLCG2: phospholipase C gamma 2; SIFD: sideroblastic anemia with B-cell immunodeficiency, periodic fevers, and developmental delay; LACC1: laccase domain containing 1; VEO-IBD: very early onset inflammatory bowel disease. Modified from (Nigrovic, Lee, and Hoffman 2020).

Autoinflammatory diseases caused by defects in the complement system are also grouped here. The complement system is a complex network within the innate immunity and is in charge of recognizing, targeting and eliminating pathogens. Genetic variants in *CFH*, *THBD*, *CFI* and *CD46* genes are associated with atypical haemolytic uremic syndrome, and others in *CFH* are associated with age-related macular degeneration (Manthiram et al. 2017).

2.3. Genetic variation in autoinflammatory diseases

In this thesis, I have already discoursed about genetic variation in humans and how to discover it through MPS experiments, the immune system and the inflammation processes in physiological and pathological conditions (i.e. autoinflammatory diseases). To close the circle, the Introduction finishes with this last section about genetic variants already described to cause autoinflammatory diseases.

The possible genetic origin of both monogenic and polygenic autoinflammatory diseases has been studied and, although genetic risk factors have been identified for the polygenic ones, pathogenic variants have been discovered only in the monogenic scenario.

Polygenic autoinflammatory diseases are harder to study, and a low number of genes with a weak effect have been linked to them. For instance, in the Behçet syndrome, HLA-B*51 has been identified as an important risk factor, but the reality is that its prevalence among patients with this disease is low (Yazici et al. 2018). In inflammatory bowel disease, more genes have been described as risk factors, such as *NOD2*, *TNFSF15*, *IL23R*, *PTPN2* and others. Genetic studies of inflammatory bowel disease have been recently reviewed (Graham and Xavier 2020).

As aforementioned, to date, more than 40 different monogenic autoinflammatory diseases and their associated genetic defects have been described (Tangye et al. 2020). The importance of germline SNVs in autoinflammatory diseases is well known from the first discoveries of the field. Four different missense causal SNVs were already described in *MEFV* gene for the FMF in the first paper of autoinflammatory diseases: c.1130G>C, c.1170A>G, c.1172G>A and c.1267T>C (Santer et al. 1997).

From that moment on, hundreds of different germline SNVs and indels, and some structural variants, have continued being described to cause autoinflammatory diseases. Just to mention some of them we can highlight the following:

Several homozygous and compound heterozygous mutations in *MKV* gene (c.60T>A, c.394G>A, c.511G>A, c.632G>A), plus one deletion spanning 19 nucleotides (c.16_34del) as a cause for mevalonate kinase deficiency (D'Osualdo et al. 2005). Several missense heterozygous SNVs in the third exon of *NLRP3* causing FCAS, NOMID/CINCA and MWS. Some of the first described variants were c.592G>A, c.1055C>T, c.1316C>T and c.1880A>G (Hoffman et al. 2001). Homozygous SNVs and deletions in *IL1RN* causing DIRA (c.156_157del, c.160C>T, c.229G>T) (Aksentijevich et al. 2009), as well as a deletion of 175 kb on chromosome 2 spanning five additional genes (Reddy et al. 2009). Heterozygous SNVs in *TMEM173* causing SAVI (c.439G>C, c.461A>G and c.463G>A) (Liu et al. 2014). Heterozygous SNVs in *NOD2* causing Blau syndrome (c.1000C>T, c.1001G>A, c.1405C>T) (Miceli-Richard et al. 2001).

Apart from their already discussed impact in cancer, somatic variants are known to play an important role in the genesis of autoinflammatory diseases. Somatic pathogenic variants have been described in the *NLRP3* gene causing MWS (c.914A>C, c.1046C>T, c.1237C>T, c.1239G>T, c.1311G>T, c.1575C>A, c.1697G>A), late-onset MWS (c.924A>T, c.1060G>A, c.1694A>G, c.1912C>G) and CINCA syndrome (c.913G>C, c.926G>T, c.1570A>T) (Mensa-Vilaró et al. 2018). One somatic variant in *NOD2* (c.1001G>A) has been identified in cases of Blau syndrome (De Inocencio et al. 2015; Mensa-Vilaro et al. 2016). The c.461A>G variant in *TMEM173* has also been found to cause SAVI in a somatic state (Liu et al. 2014). Also, one somatic SNV in *MEFV* gene (c.1955G>A) has been reported as a cause for FMF (Shinar et al. 2015). A 24 bp somatic deletion in *TNFRSF1A* (c.255_278del) was discovered in a patient with TRAPS (Rowczenio et al. 2016) and one somatic SNV in *NLRC4* (c.529A>G) was reported as the causal agent for NOMID (Kawasaki et al. 2017).

The role of somatic variants in autoinflammatory diseases (Hoffman and Broderick 2017) and, more widely, in immunological disorders (Van Horebeek, Dubois, and Goris 2019) has been reviewed. In comparison, the contribution of SVs to autoimmune disorders does not seem to be that strong (Yim et al. 2015).

Objectives

The general objective of this thesis was to discover new somatic and germline genetic variants causing autoinflammatory disease using massive parallel sequencing technologies. The specific objectives were:

- To assess the suitability of massive parallel sequencing methods to uncover somatic pathogenic variants in autoinflammatory diseases and other primary immunodeficiencies.
- To explore the somatic coding variant landscape in the context of individuals suffering from autoinflammatory diseases and other primary immunodeficiencies.
- To detect and characterize the possible pathogenic mutational events in a family with two cases of a disease compatible with the deficiency of IL-1 receptor antagonist (DIRA).
- To set up a methodology to isolate and enrich individual chromosomes 1 and sequence it with long reads for a more accurate structural variant calling.
- To explore and characterize a structural variant event encompassing the *NLRP3* gene in a family with a disease compatible with cryopyrin associated periodic syndrome (CAPS) taking advantage of the previously set up methodology.

Chapter 1

An assessment of the gene mosaicism burden in blood and its implications for immune disorders

Manuel Solís-Moruno[1,2], Anna Mensa-Vilaró[3,4], Laura Batlle-Masó[1,2], Irene Lobón[1], Núria Bonet[2], Tomàs Marquès-Bonet[1,5,6,7], Juan I. Arostegui[3,4,8], Ferran Casals[2]

Manuscript in revision in *Scientific Reports*.

Abstract

There are increasing evidences showing the contribution of somatic genetic variants to non-cancer diseases. However, their detection using massive parallel sequencing methods still has important limitations. In addition, the relative importance and dynamics of somatic variation in healthy tissues are not fully understood. We performed high-depth whole exome sequencing in 16 samples from patients with a previously determined pathogenic somatic variant for a primary immunodeficiency and tested different variant callers detection ability. Subsequently, we explored the load of somatic variants in the whole blood of these individuals and validated it by amplicon-based deep sequencing. Variant callers allowing low frequency read thresholds were able to detect most of the variants, even at very low frequencies in the tissue. The genetic load of somatic coding variants detectable in whole blood is low, ranging from 1 to 2 variants in our dataset, except for one case with 17 variants compatible with clonal haematopoiesis under genetic drift. Because of the ability we demonstrated to detect this type of genetic variation, and its relevant role in disorders such as primary immunodeficiencies, we suggest considering this model of gene mosaicism in future genetic studies and considering revisiting previous massive parallel sequencing data in patients with negative results.

Introduction

The distribution and effect of somatic genetic variants in disease has been studied mostly in cancer. However, in the past years, they have also been identified in a wide spectrum of syndromes including neurological disorders as schizophrenia (Bundo et al. 2014), autism spectrum disorder (D'Gama et al. 2015), Alzheimer (Beck et al. 2004; Bushman et al. 2015; Parcerisas et al. 2014; Sala Frigerio et al. 2015) or Huntington disease (Swami et al. 2009), coronary heart disease and stroke (Jaiswal et al. 2014) and kidney diseases such as the Alport syndrome (Bruttini et al. 2000; Krol et al. 2008; Plant et al. 2000). In fact, at least theoretically, all monogenic diseases could be originated by a postzygotic mutation and the resulting somatic mosaicism. In the field of immune-related diseases, a remarkable number of somatic variants have been described in monogenic autoinflammatory diseases (Bessler et al. 1994; Kawasaki et al. 2017; Mensa-Vilaro, Tarng Cham, et al. 2016; Mensa-Vilaro, Teresa Bosque, et al. 2016; Saito et al. 2005; Saito et al. 2008; Takeda et al. 1993; Tanaka et al. 2011; Zhou et al. 2015), and a recent work has shown its important contribution to these disorders and other primary immunodeficiencies (PIDs) (Mensa-Vilaró et al. 2018).

Understanding the relative abundance of somatic variants in health is critical to design efficient tools for mosaicism detection in disease studies. Different studies have measured the presence of somatic variation in normal tissues, most assessing the presence of mutations in cancer-driver genes, such as *NOTCH1* mutations, which undergo expansion through positive selection (Martincorena et al. 2015; 2018; Yokoyama et al. 2019). They reported the colonization of the tissue by mutant clones increasing with age and exposure to mutagenic agents (sun radiation, tobacco). Other studies, based on single cell (Lodato et al. 2015) or transcriptome analysis (García-Nieto, Morrison, and Fraser 2019) revealed tissue-specific patterns of somatic variant distribution, as well as negative selection of functional variants in non-cancer samples.

The detection of somatic variants from massive parallel sequencing (MPS) data presents some difficulties. Standard variant calling methods are based on the presence of germline heterozygous mutations in about 50% of the sequencing reads, and may fail to detect somatic variants in allelic imbalance and lower frequencies. Most of the algorithms developed for somatic variant analyses have been optimized for cancer studies where a tumour sample is compared with the healthy tissue from the same individual (Cai et al. 2016; Hofmann et al. 2017; Krøigård et al. 2016; Xu et al. 2014). Of note, studies comparing the output of different variant callers have revealed low levels of overlap (Cai et al. 2016; Krøigård et al. 2016). The tumour vs. healthy tissue approach is not suitable for somatic variant detection in mosaicisms, where the same postzygotic variant might be present in several tissues at similar frequencies. Alternatively, other variant calling tools can be applied to non-matched samples (Sandmann et al. 2017; Teer et al. 2017). In this case, allelic imbalance thresholds will need to be relaxed to detect low frequency variants, at the cost of substantially increasing the number of candidate variants. Then, an adequate filtering strategy will be essential to differentiate sequencing artefacts from true genetic variants. These filters are based both on technical criteria to exclude sequencing or mapping errors and biological knowledge to restrict the analysis to a set of candidate regions. A validation step, such as amplicon-based deep sequencing (ADS), will be ultimately required to confirm the presence of a somatic variant and better determine its frequency.

In the present study we aim to assess the load of somatic coding variants in peripheral blood at detectable frequencies from MPS data, which is relevant to detect somatic causal variants in monogenic Mendelian diseases, in particular PIDs. These diseases represent a privileged scenario for the study of the somatic pathogenic variation because of the needed presence of the causal variant in blood, as well as probably in other easily accessible tissues, and the reported important contribution of somatic mutation in these disorders (Mensa-Vilaró et al. 2018). For this, we initially performed whole-exome sequencing (WES) in a total of 16 samples belonging to 12 individuals. All individuals carry a pathogenic and previously described somatic mutation related to a PID while one patient carries a germline variant. We then selected the best candidate somatic variants, based on read quality and mapping information, to be validated with ADS. With this analysis we have tested the ability to detect causal somatic variation in PID as well as estimated the actual number of functional coding variants in blood at detectable frequencies from WES data.

Material and methods

Ethical Approval

Written informed consents for genetic analyses and participation in the study were obtained from each enrolled individual. The Ethics Committees of Hospital Clínic and Universitat Pompeu Fabra (reference number 7HCB/2019/0631), both located in Barcelona, approved the study, which was carried out in accordance with the principles and last amendments of the Declaration of Helsinki.

Samples

The present study included both unique and matched samples from peripheral blood (PB), oral mucosa (OM) and urine (UR) for 12 individuals: i) 11 unrelated PID patients carrying a pathogenic and previously described somatic variant, and ii) one of the descendants with the same pathogenic variant in germline status (Table 1). In eight individuals, the only analysed sample was PB (S2, S4a, S6, S8, S9, S10 and S11) or OM (S4). In four individuals, we analysed samples from paired tissues: from PB and OM in three patients (S1a-S1b, S3a-S3b and S5a-S5b) and, in the remaining patient, from PB and UR (S7a-S7b).

All of the PID mutations are missense single nucleotide variants (SNVs), and are the disease causing mutation either in the proband or in its offspring, where they are germline variants. The range of variant allele frequencies (VAFs) for the somatic variants previously estimated by ADS (Mensa-Vilaró et al. 2018) ranges from 2.3% to 34.8%.

For patient S5 we included additional samples from urine, oral mucosa, whole blood (before and after anti-IL-1 treatment), and different cell type populations previously isolated by flow cytometry (Mensa-Vilaro, Bosque, et al. 2016): neutrophils, monocytes, B cells, T CD4+ cells and T CD8+ cells (all pre-treatment).

Sequencing and Genomic Analysis

After DNA extraction, library preparation and exome capture were performed with the Nextera Rapid Capture kit (Illumina) according to the manufacturer's instructions. The libraries were sequenced in a NextSeq Illumina platform in three High Output 2 × 150 paired-end cycles runs to a mean coverage of 245X. We used BWA-mem version 0.7.16a-r1181 (Li 2013) to map the samples to the human reference genome hg38 (UCSC). We marked duplicated reads using Picard version 2.18.6 MarkDuplicates and realigned indels using GATK's version 3.7 (Poplin et al. 2017) IndelRealigner. We also performed base quality score recalibration using GATK's BaseRecalibrator.

We used eight publicly available tools to call genetic variants: FreeBayes version 0.9.14-8-g1618f7e (Garrison and Marth 2012), HaplotypeCaller version 3.7 (Poplin et al. 2017), LoFreq version 2.1.2 (Wilm et al. 2012), MuTect2 version 3.7 (Poplin et al. 2017), SomVarIUS version 1.1 (Smith et al. 2016), Strelka2 version 2.7.1 (Kim et al. 2018), VarDict version 1.0 (Lai et al. 2016) and VarScan2 version 2.4.3 (Koboldt et al. 2012). FreeBayes and HaplotypeCaller are purely germline callers. SomVarIUS is a caller designed to detect somatic variants in unpaired samples. The rest of them support a single mode and a paired mode. Although in our study we were not analysing cancer samples, we tested the behaviour of variant callers' paired mode in this context with the matched PB-OM and PB-UR samples. We used default parameters for all the callers except for VarScan2, where we lowered the allele frequency threshold of 20% and set the p-value to 1 to retrieve all the possible calls. For HaplotypeCaller, we first used the default ploidy parameter of 2 and next we considered other ploidy values: 4, 5, 6 and 10.

For variant calling, the manufacturer's targeted regions were intersected with our VCF files to retrieve the on target genetic variants, and we restricted our analysis to these regions. We annotated the variants using SnpEff version 4.3t (Cingolani, Platts, et al. 2012) and SnpSift version 4.3t (Cingolani, Patel, et al. 2012). Using the database dbNSFP version 4.0b1a (Liu et al. 2016), we added parameters of interest such as CADD score (Kircher et al. 2014), GERP score, ExAC (Lek et al. 2016) and gnomAD allele frequencies. We also added two functional predictions, gene haploinsufficiency values (Huang et al. 2010) and Residual Variation Intolerance Score (RVIS) (Petrovski et al. 2013).

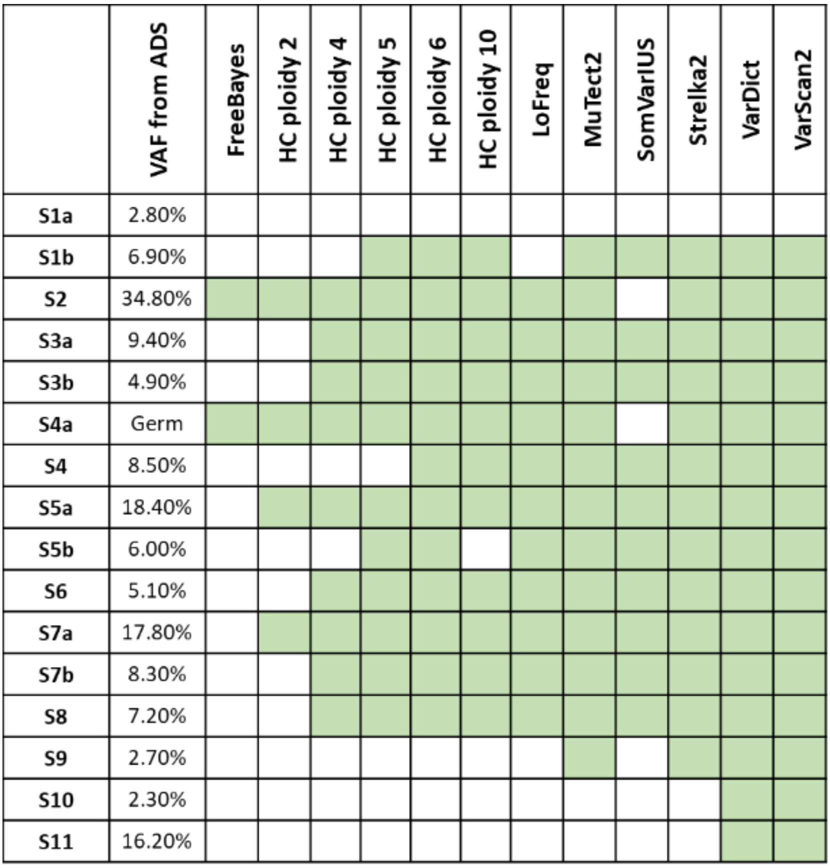
We performed ADS with rhAmpSeq from Integrated DNA Technologies (IDT, Coralville, USA) to validate the candidate somatic variants. We sequenced every selected position to a mean coverage > 20,000X in a NextSeq Illumina platform in a High Output 2×150 paired-end cycles run. The confirmed in blood plus 9 additional candidate somatic variants in S5 were analysed for validation in different tissues and cell population samples. They were sequenced in a MiSeq v3 run (2×300) to a final depth > 155,000X. We used BWA-mem version 0.7.16a-r1181 to map the fastq files to the human reference genome hg38 (UCSC). We then used pysam version 0.15.2 to count the number of reads supporting every allele, requiring a minimum mapping quality of 20 to calculate VAFs.

Results

Detection of somatic pathogenic variants from WES in PID patients

We performed WES in all DNA samples to a mean coverage of 245X (Table 1). The total number of genetic variants differs among the different callers (Supplementary Figure 1), mostly because of VarDict and VarScan2, the two callers with relaxed allelic imbalance parameters, which called more than 200,000 variants each. These two callers also show high heterogeneity across samples, which correlates with sequencing depth, as expected in MPS experiments. The amount of overlapping variants across the different callers is uneven, especially for SomVarIUS, due to the low number of variants it calls. The number of concordant variants between VarDict and VarScan2 is also low, probably because VarDict calls 3-4 times the number of indels of Varscan2 and because of discrepancies calling low frequency variants (Supplementary Figure 2).

Figure 1 shows which known causal somatic variants (Table 1) are detected by each software. FreeBayes and HaplotypeCaller have the lowest detection ratios. For the rest, the ability of detection is similar and seems to depend on the frequency of the mutations, along with the coverage of the sample and the mapping quality. The S1a causal variant has not been called by any software, but visual inspection of the mapped reads revealed that none of them supported the alternate allele (Supplementary Figure 3). Excluding it, VarDict and VarScan2 were able to detect all the causal variants. To increase the power of detection of HaplotypeCaller, we explored the effect of modifying the ploidy parameter. We used ploidy 2 (default), 4, 5, 6 and 10 in order to call variants with lower frequencies than expected in a germline scenario. This parameter is normally tuned when working with organisms with ploidies different than 2. For instance, decaploid plants have been reported (Ahmadi and Bringhurst 2019; Hummer, Nathewet, and Yanagi 2009), and genotypes 0/0/0/0/0/0/0/0/0/1 are possible. This way, the increase of the ploidy parameter makes HaplotypeCaller more sensible to low frequency variants. The percentage of detected variants increased sequentially with the ploidy parameter, although some remained undetected. HaplotypeCaller seems to be sensitive to mapping quality as in the case of the *ELANE* region (Supplementary Figure 4), where a variant with moderate frequency is not detected by this caller. Interestingly, we lost one variant using ploidy 10 while it was previously detected with ploidies 5 and 6 due to memory reasons (Figure 1, expanded in Supplementary Figure 5).



	VAF from ADS	FreeBayes	HC ploidy 2	HC ploidy 4	HC ploidy 5	HC ploidy 6	HC ploidy 10	LoFreq	MuTect2	SomVarIUS	Strelka2	VarDict	VarScan2
S1a	2.80%												
S1b	6.90%				🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩
S2	34.80%	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩
S3a	9.40%			🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
S3b	4.90%			🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
S4a	Germ	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩
S4	8.50%					🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
S5a	18.40%		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
S5b	6.00%				🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩
S6	5.10%			🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
S7a	17.80%		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
S7b	8.30%			🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
S8	7.20%			🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
S9	2.70%								🟩		🟩	🟩	🟩
S10	2.30%											🟩	🟩
S11	16.20%											🟩	🟩

Figure 1. Previously reported causal somatic mutations detected by each variant caller (in green), assessed as the presence of the variant in the raw VCF files. The germline variant in S4a was detected in Strelka germline mode but not in the somatic one. All VAF were extracted from a previous publication (Mensa-Vilaró et al. 2018).

Next, we assessed the performance of the five variant callers including a paired mode in the four cases with available paired samples (S1, S3, S5 and S7), where the same variant is present in two tissues with different frequencies. As a general trend, there is no improvement of the detection rate when using the paired mode compared to the single mode, probably because of the small differences in allele frequency between tissues. The use of one or the other paired sample as cancer/healthy tissue does not seem to affect the capacity of detection. Again, VarDict and VarScan2 showed the best detection ratios (Supplementary Table 1).

Filtering strategies for the identification of true causal variants

Once genetic variants have been called, a set of different filters is commonly applied to reduce the number of false positives. This is a crucial issue in the study of monogenic syndromes, where the aim is moving from the approximately 20,000 genetic variants identified in a typical WES to one or a few candidate variants. Relaxing or disabling the VAF filters to increase the ability to detect causal somatic variants, as we did in this study, produces an important increase of the number of mutations per individual, making this process highly recommended.

We evaluated the ability to identify the known pathogenic variants after applying the standard filters to the variants called by VarDict and VarScan2, the most successful programs in calling them (Figure 1). We started by intersecting the two VCF files for every individual, given that in all cases the true variants were retained by both of them. Next, we applied a set of additional filters sequentially (see below), checking in every step if the causal variant was retained or filtered out (Table 2). First, we filtered out SNPs located 6bp around indels. Second, as suggested previously (Bae et al. 2018), we restricted our analysis to the 1000 Genomes Project strict mask filter. Third, we required the positions to be covered by, at least, 50 reads (DP > 50) and to show a minimum quality value of 25 (QUAL > 25). Fourth, we only kept loss of function and missense variants. Fifth, we applied a stringent population allele frequency threshold of 0.001 in gnomAD. With a high probability, a somatic variant will be absent in the population because of its *de novo* nature, although the possibility of having a recurrent mutation cannot be excluded. Sixth, following the recommendations in the literature, we kept variants with a likely damaging predicted effect (CADD > 15 (Kircher et al. 2014)) and a high evolutionary conservation score, as an indicator of its functional importance (GERP > 2 (Myers et al. 2010)). Seventh, we required at least three reads supporting the alternate allele (VD ≥ 3) in every call. Finally, we used the list of 333 genes of the International Union of Immunological Societies (IUIS, updated in February 2018) (Picard et al. 2018) as a set of candidate genes for PIDs. Excluding the causal somatic variant of sample S1a, which was not detected in the sequencing process, 13 out of the 14 somatic mutations were included in the final list of candidate variants. The remaining one (S6), was filtered out because of a GERP value lower than 2.

Mosaicism abundance detection in whole blood

As mentioned above, the consideration of genetic variants deviating from the approximate expectation of 50% read frequency increases substantially the number of called variants. In the previous analyses we assessed how many of the true causal variants in 11 PID samples were detected. Now we wonder what proportion of the called variants in these samples corresponds to real postzygotic mutations, and not to sequencing, mapping or calling errors. We restricted the analysis to coding variants, more prone to have a functional impact and to be related to monogenic disorders. For this, we applied the following filters to select the variants more plausible to be validated as true: we intersected the SNPs called by VarDict and VarScan2, removed SNPs located 6 bps around indels, applied 1000G strict mask, required a minimum depth of 50 and a minimum quality of 25, removed variants classified as common in dbSNP and those shared among samples in the study, removed SNPs located within homopolymers, and removed SNPs in positions where the mappability was not perfect. We also performed a binomial test to exclude potential heterozygous mutations, to estimate the possibility of the observed number of reads supporting the alternate allele given the total number of reads. We finally required a minimum number of reads supporting the alternate allele of 7, due to the large number of variants below this threshold in our dataset (Supplementary Figure 6). After this filtering, we moved from the approximately 250,000 variants called per individual to around 40. (Figure 2), representing a total of 461 candidate somatic variants (Supplementary Table 2) for the 11 blood samples. 327 (70%) of the variants were missense, while 92 (20%) were synonymous and 19 (4%) were stop-gain. The remaining 23 variants were annotated as structural interaction variants and splice variants. Remarkably 30 of the variants were located in zinc finger proteins, 20 of them located in chromosome 19, and none of them were validated.

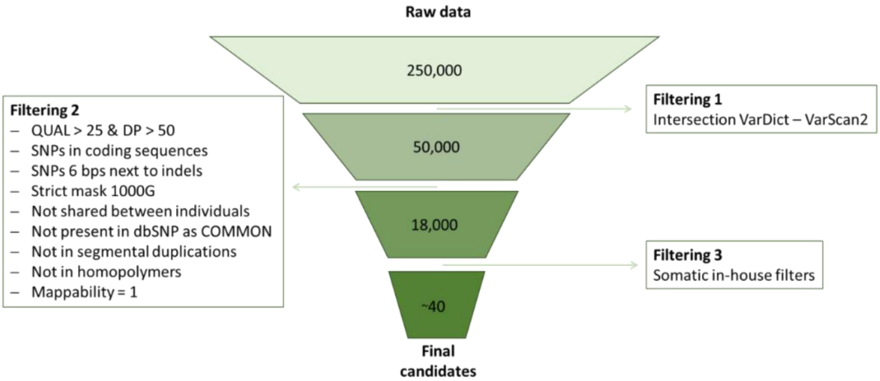


Figure 2. Filtering process followed to obtain somatic candidate variants. We got around 40 variants per blood sample that we then experimentally validated by ADS.

The 461 candidate variants were analysed by ADS with the rhAmpSeq technology (see Methods). All candidate positions were resequenced in the individual in which they were called and in the rest of individuals, plus two healthy individuals as controls. The average coverage per position was 22,500X (max=272,401, min=0, sd=21,296). The overall validation ratio was very low. For five individuals (S6, S7, S8, S9 and S10), only the initial pathogenic variant was validated, with none of the other additional candidate variants confirmed. In other six individuals, including the individual with no somatic variants (S4a), we validated one additional variant: one missense variant in *ODF2* (S1), *SHISHA2* (S2), *STRIP1* (S3) and *IL2RG* (S11), and one synonymous variant in *CACNAS1* (S4) and *ROBO4* (S11). Of note, in patient S5 we validated a total of eleven variants: seven missense, being one of them the causal variant in *NLRP3*, and four synonymous. The twelve variants seemed to cluster in two frequency groups: one with variants of about 25% (including the pathogenic variant) and other with variants about 4.5% (Supplementary Table 2).

Cell type distribution of somatic variants in S5 patient

Given the high number of validated somatic variants in patient S5, we expanded the analysis selecting nine additional candidate genetic variants. These variants were analysed for validation, along with the twelve previously confirmed, both in the whole blood sample and different cell populations separated by flow cytometry (Mensa-Vilaro, Teresa Bosque, et al. 2016) (Table 3). We also added a whole blood DNA extraction obtained after the anti-IL-1 treatment this patient received. In this experiment, the average coverage per position was 158,000X (max=484,219, min=16,689, sd=80,940). We considered that a somatic variant was validated in a given cell type or tissue when the proportion of reads supporting the alternate allele was above 0.30%, a value close to the average error type of sequencing by synthesis technologies, which also varies with features such as sequence context or the specific nucleotide change (Salk, Schmitt, and Loeb 2018; Pfeiffer et al. 2018). Six of the nine new genetic variants were validated, with one (chr7:157,614,060) being a germline variant according to its frequency (Table 3).

Overall, we detected 17 somatic variants in this patient, 16 protein coding and one intronic (Table 3), now clustered in three groups with similar VAFs around 24%, 4.5% and 1.5% in whole-blood pre-treatment (Figure 3) and cell type distribution. VAFs changes across different cell types and tissues are coordinated within each group, being the two main groups only present in the myeloid line as well as in urine and cell mucosa, but absent in the lymphoid line. In general, we found higher allele frequencies in monocytes and lower in oral mucosa. The presence of the somatic variants in oral mucosa and urine was produced by leukocyte infiltration, which was detected by flow cytometry (Mensa-Vilaro, Teresa Bosque, et al. 2016). On the other hand, the lowest VAF group of variants are detected in myeloid cells and B cells, but not in T cells. The VAF of all the somatic variants is reduced in the whole-blood sample after the anti-IL-1 treatment (Whole blood 3 post, in Table 3). This decrease is more important for the variants restricted to the myeloid line, and it is likely observed because of the increased proliferation of inflammatory cells, which is now controlled with the treatment (Mensa-Vilaro, Teresa Bosque, et al. 2016).

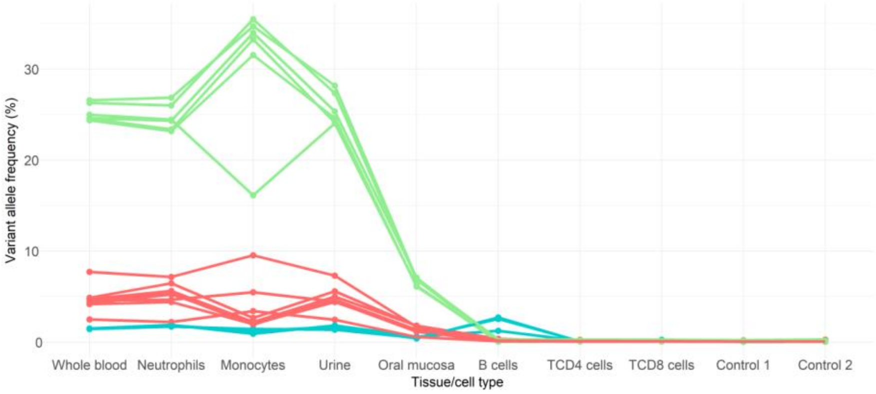


Figure 3. VAF of validated somatic variants in S5 patient per tissue and cell type. Green is used for the group of variants with higher VAF (around 24%), red for those with intermediate VAF (around 4%) and blue for those with low VAF (around 1.5%, the only group present in B cells). Of note, there is one variant in the X chromosome whose frequency has been divided by 2 in order to visualize it grouped with the others.

Discussion

We performed WES of DNA samples from patients with PIDs, carrying variable degree of gene mosaicism and assessed the ability to detect the somatic causal genetic variants by using different tools. Among the eight variant callers tested, VarDict and VarScan showed the higher detection rates of the causal somatic variants. The rest of the callers designed for somatic variant detection (MuTect2, SomVarIUS and Strelka2) mainly showed some limitations with the lower frequency variants at lower coverage. FreeBayes and HaplotypeCaller, designed for germline variant detection, failed to detect most of the somatic mutations. However, the performance of HaplotypeCaller increases when modifying the ploidy parameter, devised for non-diploid organisms and which allowed retrieving variants with less frequency than the expected 50% in the germline. Of interest, the efficiency of the five callers including a paired mode did not increase when using paired samples, probably because of the small frequency difference between the two samples carrying the same mutation.

Allele frequency is the main limitation for calling a somatic variant, with the risk of non-capturing the mutation because of its low frequency and/or insufficient coverage. To capture these low frequency variants, sequencing depths should ideally be higher than the commonly average depths achieved in WES studies (60-100X). However, the average coverage value might not be informative enough on the sequencing performance for all genomic regions, given the non-uniformity of the capture process. The use of new metrics including this information has been proposed (Wang et al. 2017), which should help to reduce false-negative results. As an example, the *NOD2* region is clearly captured more efficiently than the *NLRP3* region in our study (Table 1). On the other side, only a few reads supporting the alternate allele seems enough to detect the variant, with as few as 3 (out of 128) for the S10 variant or 7 (out of 97) for the S1b variant (Table 1). Thus, an increase of the sequencing depth to 100-200X is recommendable in cases in which somatic variation is suspected. Higher coverage facilitates the detection of very low frequency variants, but increases the risk of enlarging the list of candidate variants because of approaching the error rate of MPS technologies (Schirmer et al. 2015).

Genetic studies usually implement a set of filters to reduce the number of candidate variants to the causal one or to a small group. This process is a trade-off between reducing the number of false positives (either sequencing or mapping artefacts, and non-causal variants) and false negatives (called but filtered true causal variants). At the risk of missing the causal variant, these filters are essential to determine, at least, a reduced list of candidate genes for monogenic syndromes. In the case of studies like this, where the relaxation of allele frequency thresholds generates a list of up to hundreds of thousands of variants per sample (Supplementary Figure 1), this step can be especially critical. After applying commonly used filtering parameters both for sequencing and biological features, only the causal variant in one patient was discarded because of low conservation score (GERP for S6 causal variant: -8.07). In the case of applying more stringent filters, two more variants (S1b and S5a-b) would be missed due to GERP score vale lower than 4 (Amendola et al. 2015; de Valles-

Ibáñez et al. 2016). On the other side, only S6 causal variant would not pass a CADD threshold of 20.

The final number of candidate genetic variants exceeds by about ten times the number of variants in studies analysing germline variants. Considering the IUIS list of 333 candidate genes for PIDs, this is still quite high, with approximately 0.5 variants per gene in each individual. Therefore, it seems recommendable to restrict the analysis to a reduced set of candidate genes according to the clinical phenotype of each patient. Alternatively, the use of some gene features could also help to reduce the list of candidate variants if there is not any *a priori* clear candidate. Several gene indexes have been developed to measure their possible contribution to human disease. Among them, haploinsufficiency predictions could seem useful for identifying candidate genes in a somatic variant disease model expecting to follow a dominant inheritance pattern. However, all the genes with somatic causal variants included in this study show haploinsufficiency values below the consensus threshold of 0.5, with *NLRP3*, a gene that is proven to be mutated in different autoinflammatory diseases (de Torre-Minguela, del Castillo, and Pelegrín 2017), showing the highest value of 0.465. In contrast, *NLRP3* has been reported as a gene with a high level of intolerance to functional variation (RVIS=-0.95, in the top 9.38% of genes) (Mensa-Vilaró et al. 2018).

It is important to consider that exome sequencing was performed in DNA samples obtained from peripheral blood. Therefore, only somatic variants present in the major cell populations in blood can be detected. Neutrophils represent more than half of the nucleated blood cells (55%-75%) in healthy individuals, while lymphocytes represent around 20% (from which T cells are ~70%, B cells are just ~20%, and NK cells ~10%) (Berrington et al. 2005). Thus, for early postzygotic mutations, the capacity of detection will most probably not be affected by the cell type implicated in the disorder, since the variant will have similar frequencies in all cell populations. In contrast, for later onset mutations restricted to particular lineages, the mutation will only be detectable if present in the major cell populations of the analysed tissue.

Therefore, for immune disorders, the probability of detecting a causal variant from whole-blood extraction analysis will be much higher in those produced by alteration in the myeloid cells, such as in autoinflammatory disorders, than in the lymphoid cells. This fact can partially explain the larger number of reported cases in autoinflammatory disorders (Mensa-Vilaró et al. 2018) compared to other PIDs, as well as the lack of success in the identification of somatic variants in lymphoid immunodeficiencies such as CVID (de Valles-Ibáñez et al. 2018). In these latter situations, it is expected that a big proportion of somatic causal variants would only be detectable if the analysis is restricted to particular cell types. Thus, cell subsets isolation can be essential to the identification and/or the validation of somatic genetic variants in these less represented cell types.

Beyond the detection of the known causal variants, the detected load of coding variants per exome was very low. Except for S5, all the individuals carry none or only one somatic variant additionally to the causal variant. The vast majority of candidate variants were false positives, even if they passed the mapping and quality filters. Comparing our results to other studies is not straightforward because of the differences in the methodologies used and the scanned VAFs, as well as the conceptual approach and targeted regions (see Introduction). A whole-genome sequencing (WGS) data analysis of 11,262 blood samples revealed a median number of three mosaic mutations for younger individuals, increasing after 35-45 years of age, and considering 20 somatic variants as the threshold for clonal expansion, that affecting 12.5% of the individuals (Zink et al. 2017). Although the minimum detectable VAF of the study was limited because of the 34.8X mean coverage, the results seem concordant with the low number of somatic variants described in our WES deep sequencing approach. In addition to scanning a wide range of VAFs, we validated our results by ADS, which confirmed the low number of somatic coding variants detectable in blood. At a finer level, the total number of somatic variants per cell has been estimated in single-cell studies (García-Nieto, Morrison, and Fraser 2019; Lodato et al. 2015), although most of this variation would remain undetected when the whole tissue is analysed. In fact, when much lower frequencies have been scanned (VAF≥0.0001), it has been shown that clonal haematopoiesis is present in up to 97% of middle-aged people (Young et al. 2019). However, in absence of positive selection on a given mutation, only those that occurred earlier would reach detectable frequencies.

We identified a particular patient with an excess of validated variants compared to the others. S5 is the oldest individual of our dataset (64 years old), although another individual of similar age was also included in this study. Especially for the higher VAF group of five variants (which includes the causal one in *NLRP3*), the frequency pattern is quite uniform, except for one of the variants in chromosome X (chrX:71,537,899), with lower frequency in monocytes. The presence of the genetic variants in the lowest frequency cluster in cells of the myeloid lineage and in B cells, but not in T cells, could be explained by its origin in adult hematopoietic stem cells generating multineage outputs (Lee-Six et al. 2018). Because of the seemingly aggrupation in three different clusters of frequencies and cell type distribution, we propose simultaneous occurrence and clonal expansion as the most parsimonious explanation. However, none of the genes with somatic variants in S5 (Table 3) seems to be related with cellular proliferation that could be linked to an adaptive advantage of a clone of cells, and we also discarded the presence of additional candidate variant in *DNMT3A*, *TET2* and *ASXL1* genes, known to be implicated in hematologic malignancies (Genovese et al. 2014; Jaiswal et al. 2014). In fact, in the aforementioned study of WGS of 11,262 individuals (Zink et al. 2017) only 12.6% of the cases of clonal haematopoiesis had detectable cancer driver mutations. Thus, on the rest of cases as well as for S5, clonal haematopoiesis could be produced by genetic drift, as suggested in simulation analysis (Klein and Simons 2011). In contrast, a recent study (Watson et al. 2020) proposes positive selection being the major driving force of clonal haematopoiesis, and that it would take more than 2,000 years for a mutation to reach a VAF > 1% by only drift. However, our results do not seem to fit to this explanation, because of the abovementioned gene location as well as the presence of synonymous and intronic variants.

Finally, although we believe that our study contributes to the understanding of the burden of functional somatic mutations in blood and provides some practical advice on its detection, we would like to acknowledge some limitations of our approach. Allele frequency and sequencing depth are the two main limiting factors to detect a somatic variant as shown in our case by the failure to detect a variant with VAF<3%. Also, the number of genetic variants depends on the selected software, that show a limited level of overlapping among them. In this sense, we recommend an inclusive strategy by using the less stringent callers or parameters, followed by a filtering strategy based on sequencing and mapping features. However, even by using stringent filters, the capacity of detection of causal variants will be mostly limited to previously known candidate or related genes, given the excessive number of variants when considering the whole exome. Gene functional relevance or mutation tolerance indexes could be used to reduce the number of candidate genes, but they also show limited applicability. Of importance, we also acknowledge the limitations derived from the small size of our cohort which, while allowing the study of somatic variant discovery, makes it difficult to draw conclusions in terms of dynamics of somatic variation.

Conclusions

The detectable genetic load of somatic coding variants in blood is low. A moderate increase of the commonly achieved depths in exome sequencing analyses can be enough to detect most of these variants at frequencies above the technology error rate, for which we recommend using variant callers sensitive to low VAF. Of importance, the high proportion of false positives makes mandatory their validation which will also provide a better estimation of the VAF. Given both the feasibility of this approach and the reported contribution of gene mosaicism to PIDs (Mensa-Vilaró et al. 2018), we think that this model should be considered in future sequencing studies. It can be of special interest for those disorders related to major cell populations in blood, such as autoinflammatory diseases. We also suggest reanalysing data of undiagnosed patients, especially those where the inheritance pattern in the pedigree and/or the clinical features of the patient might fit this model. Because of the high number of possible somatic variants called per individual, even after applying stringent filters, it is advisable to restrict the analysis to a set of candidate genes defined according to the clinical phenotype. Finally, our results are in agreement with the existence of clonal haematopoiesis produced by drift, and that can be related to non-cancer disorders.

Data availability

The datasets generated during and analysed during the current study are available in the European Nucleotide Archive (ENA) repository under accession code PRJEB44742.

Bibliography

Ahmadi, Hamid, and Royce S. Bringhurst. 2019. "Breeding Strawberries at the Decaploid Level." *Journal of the American Society for Horticultural Science* 117 (5): 856–62. https://doi.org/10.21273/jashs.117.5.856.

Amendola, Laura M., Michael O. Dorschner, Peggy D. Robertson, Joseph S. Salama, Ragan Hart, Brian H. Shirts, Mitzi L. Murray, et al. 2015. "Actionable Exomic Incidental Findings in 6503 Participants: Challenges of Variant Classification." *Genome Research* 25 (3): 305–15. https://doi.org/10.1101/gr.183483.114.

Bae, Taejeong, Livia Tomasini, Jessica Mariani, Bo Zhou, Tanmoy Roychowdhury, Daniel Franjic, Mihovil Pletikos, et al. 2018. "Different Mutational Rates and Mechanisms in Human Cells at Pregastrulation and Neurogenesis." *Science* 555 (February): 550–55. https://doi.org/10.1126/science.aan8690.

Beck, Jonathan A., Mark Poulter, Tracy A. Campbell, James B. Uphill, Gary Adamson, Jennian F. Geddes, Tamas Revesz, et al. 2004. "Somatic and Germline Mosaicism in Sporadic Early-Onset Alzheimer's Disease." *Human Molecular Genetics* 13 (12): 1219–24. https://doi.org/10.1093/hmg/ddh134.

Berrington, J. E., D. Barge, A. C. Fenton, A. J. Cant, and G. P. Spickett. 2005. "Lymphocyte Subsets in Term and Significantly Preterm UK Infants in the First Year of Life Analysed by Single Platform Flow Cytometry." *Clinical and Experimental Immunology* 140 (2): 289–92. https://doi.org/10.1111/j.1365-2249.2005.02767.x.

Bessler, Monica, Philip J. Mason, Peter Hillmen, Toshio Miyata, Norio Yamada, Junji Takeda, Lucio Luzzatto, and Taroh Kinoshita. 1994. "Paroxysmal Nocturnal Haemoglobinuria (PNH) Is Caused by Somatic Mutations in the PIG-A Gene." *The EMBO Journal* 13: 110–17. https://doi.org/10.1016/S0140-6736(87)91654-0.

Bruttini, Mirella, Francesca Vitelli, Ilaria Meloni, Giuseppe Rizzari, Mario Della Volpe, Gianna Mazzucco, Mario De Marchi, and Alessndra Renieri. 2000. "Mosaicism in Alport Syndrome and Genetic Counseling." *Journal of Medical Genetics* 37: 717–19. http://jmg.bmj.com/content/jmedgenet/37/9/717.full.pdf.

Bundo, Miki, Manabu Toyoshima, Yohei Okada, Wado Akamatsu, Junko Ueda, Taeko Nemoto-Miyauchi, Fumiko Sunaga, et al. 2014. "Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia." *Neuron* 81 (2): 306–13. https://doi.org/10.1016/j.neuron.2013.10.053.

Bushman, Diane M., Gwendolyn E. Kaeser, Benjamin Siddoway, Jurgen W. Westra, Richard R. Rivera, Stevens K. Rehen, Yun C. Yung, and Jerold Chun. 2015. "Genomic Mosaicism with Increased Amyloid Precursor Protein (APP) Gene Copy Number in Single Neurons from Sporadic Alzheimer's Disease Brains." *ELife* 2015 (4): 1–26. https://doi.org/10.7554/eLife.05116.001.

Cai, Lei, Wei Yuan, Zhou Zhang, Lin He, and Kuo Chen Chou. 2016. "In-Depth Comparison of Somatic Point Mutation Callers Based on Different Tumor next-Generation Sequencing Depth Data." *Scientific Reports* 6 (November): 1–9. https://doi.org/10.1038/srep36540.

Cingolani, Pablo, Viral M. Patel, Melissa Coon, Tung Nguyen, Susan J. Land, Douglas M. Ruden, and Xiangyi Lu. 2012. "Using Drosophila Melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift." *Frontiers in Genetics* 3 (MAR). https://doi.org/10.3389/fgene.2012.00035.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff." *Fly* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

D'Gama, Alissa M., Sirisha Pochareddy, Mingfeng Li, Saumya S. Jamuar, Rachel E. Reiff, Anh Thu N. Lam, Nenad Sestan, and Christopher A. Walsh. 2015. "Targeted DNA Sequencing from Autism Spectrum Disorder Brains Implicates Multiple Genetic Mechanisms." *Neuron* 88 (5): 910–17. https://doi.org/10.1016/j.neuron.2015.11.009.

García-Nieto, Pablo E., Ashby J. Morrison, and Hunter B. Fraser. 2019. "The Somatic Mutation Landscape of the Human Body." *Genome Biology* 20 (1): 1–20. https://doi.org/10.1186/s13059-019-1919-5.

Garrison, Erik, and Gabor Marth. 2012. "Haplotype-Based Variant Detection from Short-Read Sequencing." *ArXiv Preprint ArXiv:1207.3907 [q-Bio.GN]*. https://doi.org/arXiv:1207.3907 [q-bio.GN].

Genovese, Giulio, Anna K. Kähler, Robert E. Handsaker, Johan Lindberg, Samuel A. Rose, Samuel F. Bakhoum, Kimberly Chambert, et al. 2014. "Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence." *New England Journal of Medicine* 371 (26): 2477–87. https://doi.org/10.1056/NEJMoa1409405.

Hofmann, Ariane L., Jonas Behr, Jochen Singer, Jack Kuipers, Christian Beisel, Peter Schraml, Holger Moch, and Niko Beerenwinkel. 2017. "Detailed Simulation of Cancer Exome Sequencing Data Reveals Differences and Common Limitations of Variant Callers." *BMC Bioinformatics* 18 (1): 1–15. https://doi.org/10.1186/s12859-016-1417-7.

Huang, Ni, Insuk Lee, Edward M. Marcotte, and Matthew E. Hurles. 2010. "Characterising and Predicting Haploinsufficiency in the Human Genome." *PLoS Genetics* 6 (10): 1–11. https://doi.org/10.1371/journal.pgen.1001154.

Hummer, Kim E, Preeda Nathewet, and Tomohiro Yanagi. 2009. "Decaploidy in Fragaria Iturupensis (Rosaceae)." *American Journal of Botany* 96 (3): 713–16. https://doi.org/10.3732/ajb.0800285.

Jaiswal, Siddhartha, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V. Grauman, Brenton G. Mar, R. Coleman Lindsley, et al. 2014. "Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes." *New England Journal of Medicine* 371 (26): 2488–98. https://doi.org/10.1056/NEJMoa1408617.

Kawasaki, Yuri, Hirotsugu Oda, Jun Ito, Akira Niwa, Takayuki Tanaka, Atsushi Hijikata, Ryosuke Seki, et al. 2017. "Identification of a High-Frequency Somatic NLRC4 Mutation as a Cause of Autoinflammation by Pluripotent Cell–Based Phenotype Dissection." *Arthritis and Rheumatology* 69 (2): 447–59. https://doi.org/10.1002/art.39960.

Kim, Sangtae, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, et al. 2018. "Strelka2: Fast and Accurate Calling of Germline and Somatic Variants." *Nature Methods* 15 (8): 591–94. https://doi.org/10.1038/s41592-018-0051-x.

Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15. https://doi.org/10.1038/ng.2892.

Klein, Allon M., and Benjamin D. Simons. 2011. "Universal Patterns of Stem Cell Fate in Cycling Adult Tissues." *Development* 138 (15): 3103–11. https://doi.org/10.1242/dev.060103.

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." *Genome Research* 22 (3): 568–76. https://doi.org/10.1101/gr.129684.111.

Krøigård, Anne Bruun, Mads Thomassen, Anne Vibeke Lænkholm, Torben A. Kruse, and Martin Jakob Larsen. 2016. "Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data." *PLoS ONE* 11 (3): 1–15. https://doi.org/10.1371/journal.pone.0151664.

Krol, Rafal Przybyslaw, Kandai Nozu, Koichi Nakanishi, Kazumoto Iijima, Yasuhiro Takeshima, Xue Jun Fu, Yoshimi Nozu, et al. 2008. "Somatic Mosaicism for a Mutation of the COL4A5 Gene Is a Cause of Mild Phenotype Male Alport Syndrome." *Nephrology Dialysis Transplantation* 23 (8): 2525–30. https://doi.org/10.1093/ndt/gfn005.

Lai, Zhongwu, Aleksandra Markovets, Miika Ahdesmaki, Brad Chapman, Oliver Hofmann, Robert Mcewen, Justin Johnson, Brian Dougherty, J. Carl Barrett, and Jonathan R. Dry. 2016. "VarDict: A Novel and Versatile Variant Caller for next-Generation Sequencing in Cancer Research." *Nucleic Acids Research* 44 (11): 1–11. https://doi.org/10.1093/nar/gkw227.

Lee-Six, Henry, Nina Friesgaard Øbro, Mairi S. Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J. Osborne, et al. 2018. "Population Dynamics of Normal Human Blood Inferred from Somatic Mutations." *Nature* 561 (7724): 473–78. https://doi.org/10.1038/s41586-018-0497-0.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285–91. https://doi.org/10.1038/nature19057.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *ArXiv:1303.3997 [q-Bio.GN]*, March. http://arxiv.org/abs/1303.3997.

Liu, Xiaoming, Chunlei Wu, Chang Li, and Eric Boerwinkle. 2016. "DbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs." *Human Mutation* 37 (3): 235–41. https://doi.org/10.1002/humu.22932.

Lodato, Michael A., Mollie B. Woodworth, Semin Lee, Gilad D. Evrony, Bhaven K. Mehta, Amir Karger, Soohyun Lee, et al. 2015. "Somatic Mutation in Single Human Neurons Tracks Developmental and Transcriptional History." *Science* 350 (6256): 94–98.

https://doi.org/10.1126/science.aab1785.

Martincorena, Iñigo, Joanna C. Fowler, Agnieszka Wabik, Andrew R.J. Lawson, Federico Abascal, Michael W.J. Hall, Alex Cagan, et al. 2018. "Somatic Mutant Clones Colonize the Human Esophagus with Age." *Science* 362 (6417): 911–17. https://doi.org/10.1126/science.aau3879.

Martincorena, Iñigo, Amit Roshan, Moritz Gerstung, Peter Ellis, Peter Van Loo, Stuart McLaren, David C. Wedge, et al. 2015. "High Burden and Pervasive Positive Selection of Somatic Mutations in Normal Human Skin." *Science* 348 (6237): 880–86. https://doi.org/10.1126/science.aaa6806.

Mensa-Vilaró, Anna, María Bravo García-Morato, Oscar de la Calle-Martin, Clara Franco-Jarava, María Teresa Martínez-Saavedra, Luis I. González-Granado, Eva González-Roca, et al. 2018. "Unexpected Relevant Role of Gene Mosaicism in Primary Immunodeficiency Diseases." *Journal of Allergy and Clinical Immunology*, 1–10. https://doi.org/10.1016/j.jaci.2018.09.009.

Mensa-Vilaro, Anna, Weng Tarng Cham, Swee Ping Tang, Sern Chin Lim, Eva González-Roca, Estibaliz Ruiz-Ortiz, Roziana Ariffin, Jordi Yagüe, and Juan I. Aróstegui. 2016. "First Identification of Intrafamilial Recurrence of Blau Syndrome Due to Gonosomal NOD2 Mosaicism." *Arthritis and Rheumatology* 68 (4): 1039–44. https://doi.org/10.1002/art.39519.

Mensa-Vilaro, Anna, María Teresa Bosque, Giuliana Magri, Yoshitaka Honda, Helios Martínez-Banaclocha, Marta Casorran-Berges, Jordi Sintes, et al. 2016. "Brief Report: Late-Onset Cryopyrin-Associated Periodic Syndrome Due to Myeloid-Restricted Somatic NLRP3 Mosaicism." *Arthritis and Rheumatology* 68 (12): 3035–41. https://doi.org/10.1002/art.39770.

Myers, R. M., A. Sidow, D. L. Goode, G. M. Cooper, J. Schmutz, M. Dickson, E. Gonzales, et al. 2010. "Evolutionary Constraint Facilitates Interpretation of Genetic Variation in Resequenced Human Genomes." *Genome Research*, no. 630: 301–10. https://doi.org/10.1101/gr.102210.109.

Parcerisas, Antoni, Sara E. Rubio, Ashraf Muhaisen, Alberto Gómez-Ramos, Lluís Pujadas, Montserrat Puiggros, Daniela Rossi, et al. 2014. "Somatic Signature of Brain-Specific Single Nucleotide Variations in Sporadic Alzheimer's Disease." *Journal of Alzheimer's Disease* 42 (4): 1357–82. https://doi.org/10.3233/JAD-140891.

Petrovski, Slavé, Quanli Wang, Erin L. Heinzen, Andrew S. Allen, and David B. Goldstein. 2013. "Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes." *PLoS Genetics* 9 (8). https://doi.org/10.1371/journal.pgen.1003709.

Pfeiffer, Franziska, Carsten Gröber, Michael Blank, Kristian Händler, Marc Beyer, Joachim L. Schultze, and Günter Mayer. 2018. "Systematic Evaluation of Error Rates and Causes in Short Samples in Next-Generation Sequencing." *Scientific Reports* 8 (1): 1–14. https://doi.org/10.1038/s41598-018-29325-6.

Picard, Capucine, H. Bobby Gaspar, Waleed Al-Herz, Aziz Bousfiha, Jean Laurent Casanova, Talal Chatila, Yanick J. Crow, et al. 2018. "International Union of Immunological Societies: 2017 Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity." *Journal of Clinical Immunology* 38 (1): 96–128. https://doi.org/10.1007/s10875-017-0464-9.

Plant, Kate E, Eileen Boye, Peter M Green, David Vetrie, and Frances A Flinter. 2000. "Somatic Mosaicism Associated with a Mild Alport Syndrome Phenotype." *Journal of Medical Genetics* 37: 238–39.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J Fennell, Mauricio O Carneiro, Geraldine A. Van der Auwera, David E Kling, et al. 2017. "Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples." *BioRxiv*, 201178.

https://doi.org/10.1101/201178.

Saito, Megumu, Akihiro Fujisawa, Ryuta Nishikomori, Naotomo Kambe, Mami Nakata-Hizume, Momoko Yoshimoto, Katsuyuki Ohmori, et al. 2005. "Somatic Mosaicism of CIAS1 in a Patient with Chronic Infantile Neurologic, Cutaneous, Articular Syndrome." *Arthritis and Rheumatism* 52 (11): 3579–85. https://doi.org/10.1002/art.21404.

Saito, Megumu, Ryuta Nishikomori, Naotomo Kambe, Akihiro Fujisawa, Hideaki Tanizaki, Kyoko Takeichi, Tomoyuki Imagawa, et al. 2008. "Disease-Associated CIAS1 Mutations Induce Monocyte Death, Revealing Low-Level Mosaicism in Mutation-Negative Cryopyrin-Associated Periodic Syndrome Patients." *Blood* 111 (4): 2132–41. https://doi.org/10.1182/blood-2007-06-094201.

Sala Frigerio, Carlo, Pierre Lau, Claire Troakes, Vincent Deramecourt, Patrick Gele, Peter Van Loo, Thierry Voet, and Bart De Strooper. 2015. "On the Identification of Low Allele Frequency Mosaic Mutations in the Brains of Alzheimer's Disease Patients." *Alzheimer's and Dementia* 11 (11): 1265–76. https://doi.org/10.1016/j.jalz.2015.02.007.

Salk, Jesse J., Michael W. Schmitt, and Lawrence A. Loeb. 2018. "Enhancing the Accuracy of Next-Generation Sequencing for Detecting Rare and Subclonal Mutations." *Nature Reviews Genetics* 19 (5): 269–85. https://doi.org/10.1038/nrg.2017.117.

Sandmann, Sarah, Aniek O. De Graaf, Mohsen Karimi, Bert A. Van Der Reijden, Eva Hellström-Lindberg, Joop H. Jansen, and Martin Dugas. 2017. "Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data." *Scientific Reports* 7: 1–12. https://doi.org/10.1038/srep43169.

Schirmer, Melanie, Umer Z. Ijaz, Rosalinda D'Amore, Neil Hall, William T. Sloan, and Christopher Quince. 2015. "Insight into Biases and Sequencing Errors for Amplicon Sequencing with the Illumina MiSeq Platform." *Nucleic Acids Research* 43 (6). https://doi.org/10.1093/nar/gku1341.

Smith, Kyle S., Vinod K. Yadav, Shanshan Pei, Daniel A. Pollyea, Craig T. Jordan, and Subhajyoti De. 2016. "SomVarIUS: Somatic Variant Identification from Unpaired Tissue Samples." *Bioinformatics* 32 (6): 808–13. https://doi.org/10.1093/bioinformatics/btv685.

Swami, Meera, Audrey E. Hendricks, Tammy Gillis, Tiffany Massood, Jayalakshmi Mysore, Richard H. Myers, and Vanessa C. Wheeler. 2009. "Somatic Expansion of the Huntington's Disease CAG Repeat in the Brain Is Associated with an Earlier Age of Disease Onset." *Human Molecular Genetics* 18 (16): 3039–47. https://doi.org/10.1093/hmg/ddp242.

Takeda, Junji, Toshio Miyata, Kazuyoshi Kawagoe, Yoshiyasu Iida, Yuichi Endo, Teizo Fujita, Minoru Takahashi, Teruo Kitani, and Taroh Kinoshita. 1993. "Deficiency of the GPI Anchor Caused by a Somatic Mutation of the PIG-A Gene in Paroxysmal Nocturnal Hemoglobinuria." *Cell* 73 (4): 703–11. https://doi.org/10.1016/0092-8674(93)90250-T.

Tanaka, Naoko, Kazushi Izawa, Megumu K. Saito, Mio Sakuma, Koichi Oshima, Osamu Ohara, Ryuta Nishikomori, et al. 2011. "High Incidence of NLRP3 Somatic Mosaicism in Patients with Chronic Infantile Neurologic, Cutaneous, Articular Syndrome: Results of an International Multicenter Collaborative Study." *Arthritis and Rheumatism* 63 (11): 3625–32. https://doi.org/10.1002/art.30512.

Teer, Jamie K., Yonghong Zhang, Lu Chen, Eric A. Welsh, W. Douglas Cress, Steven A. Eschrich, and Anders E. Berglund. 2017. "Evaluating Somatic Tumor Mutation Detection without Matched Normal Samples." *Human Genomics* 11 (1): 1–13. https://doi.org/10.1186/s40246-017-0118-2.

Torre-Minguela, Carlos de, Pablo Mesa del Castillo, and Pablo Pelegrín. 2017. "The NLRP3 and Pyrin Inflammasomes: Implications in the Pathophysiology of Autoinflammatory Diseases." *Frontiers in Immunology* 8 (JAN). https://doi.org/10.3389/fimmu.2017.00043.

Valles-Ibáñez, Guillem de, Ana Esteve-Solé, Mònica Piquer, E. Azucena González-Navarro, Jessica Hernandez-Rodriguez, Hafid Laayouni, Eva González-Roca, et al. 2018. "Evaluating the Genetics of Common Variable Immunodeficiency: Monogenetic Model and Beyond." *Frontiers in Immunology* 9 (MAY): 1–15. https://doi.org/10.3389/fimmu.2018.00636.

Valles-Ibáñez, Guillem de, Jessica Hernandez-Rodriguez, Javier Prado-Martinez, Pierre Luisi, Tomàs Marquès-Bonet, and Ferran Casals. 2016. "Genetic Load of Loss-of-Function Polymorphic Variants in Great Apes." *Genome Biology and Evolution* 8 (3): 871–77. https://doi.org/10.1093/gbe/evw040.

Wang, Qingyu, Cooduvalli S. Shashikant, Matthew Jensen, Naomi S. Altman, and Santhosh Girirajan. 2017. "Novel Metrics to Measure Coverage in Whole Exome Sequencing Datasets Reveal Local and Global Non-Uniformity." *Scientific Reports* 7 (1): 1–11. https://doi.org/10.1038/s41598-017-01005-x.

Watson, Caroline J, A L Papula, Gladys Y P Poon, Wing H Wong, and Andrew L Young. 2020. "The Evolutionary Dynamics and Fitness Landscape of Clonal Hematopoiesis." *Science* 1454 (March): 1449–54.

Wilm, Andreas, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. 2012. "LoFreq: A Sequence-Quality Aware, Ultra-Sensitive Variant Caller for Uncovering Cell-Population Heterogeneity from High-Throughput Sequencing Datasets." *Nucleic Acids Research* 40 (22): 11189–201. https://doi.org/10.1093/nar/gks918.

Xu, Huilei, John DiCarlo, Ravi V. Satya, Quan Peng, and Yexun Wang. 2014. "Comparison of Somatic Mutation Calling Methods in Amplicon and Whole Exome Sequence Data." *BMC Genomics* 15 (1): 1–10. https://doi.org/10.1186/1471-2164-15-244.

Yokoyama, Akira, Nobuyuki Kakiuchi, Tetsuichi Yoshizato, Yasuhito Nannya, Hiromichi Suzuki, Yasuhide Takeuchi, Yusuke Shiozawa, et al. 2019. "Age-Related Remodelling of Oesophageal Epithelia by Mutated Cancer Drivers." *Nature* 565 (7739): 312–17. https://doi.org/10.1038/s41586-018-0811-x.

Young, Andrew L, R. Spencer Tong, Brenda M Birmann, and Todd E Druley. 2019. "Clonal Hematopoiesis and Risk of Acute Myeloid Leukemia." *Haematologica* 104 (12). https://doi.org/10.3324/haematol.2018.215269.

Zhou, Qing, Ivona Aksentijevich, Geryl M. Wood, Avram D. Walts, Patrycja Hoffmann, Elaine F. Remmers, Daniel L. Kastner, and Amanda K. Ombrello. 2015. "Cryopyrin-Associated Periodic Syndrome Caused by a Myeloid-Restricted Somatic NLRP3 Mutation." *Arthritis and Rheumatology* 67 (9): 2482–86. https://doi.org/10.1002/art.39190.

Zink, Florian, Simon N. Stacey, Gudmundur L. Norddahl, Michael L. Frigge, Olafur T. Magnusson, Ingileif Jonsdottir, Thorgeir E. Thorgeirsson, et al. 2017. "Clonal Hematopoiesis, with and without Candidate Driver Mutations, Is Common in the Elderly." *Blood* 130 (6): 742–52. https://doi.org/10.1182/blood-2017-02-769869

Acknowledgments

This study was funded by grant SAF2015-68472-C2-2-R from the Ministerio de Economía y Competitividad (Spain) and by Direcció General de Recerca, Generalitat de Catalunya (2017SGR-702) to F.C. M.S.-M. is supported by the Ministerio de Economía y Competitividad, Spain (Maria de Maetzu grant MDM-2014-0370-16-3). L.B.-M. is supported by a Formació de personal Investigador fellowship from Generalitat de Catalunya (2018_FI_B00072). T.M-B. is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). Supported in part by CERCA Programme/Generalitat de Catalunya (JI.A.), SAF2015-68472-C2-1-R grant from the Ministerio de Economía y Competitividad (Spain) co-financed by European Regional Development Fund (ERDF) (JI.A.), RTI2018-096824-B-C21 grant from the Ministerio de Ciencia, Innovación y Universidades (Spain) co-financed by ERDF (JI.A.), AC15/00027 grant from the Instituto de Salud Carlos III / Transnational Research Projects on Rare Diseases (JI.A.).

Author contributions

F.C., JI.A, T.M.-B and M.S.-M. conceived and designed the study. M.S.-M., A.M.-V., L.B.-M. and I.L. analysed data. M.S.-M., A.M.-V. and N.B. performed laboratory work. All authors participated in the writing and correction of the manuscript.

Competing interests

The authors declare no competing interests.

Table 1. Samples and mutations included in the study.

Sample	Coordinate (hg38)	Gene	Change in DNA	Change in protein	WES			ADS
					VAF (%)	DP/VD	Mean coverage	VAF (%)
S1a (PB)	chr1:247,424,492	*NLRP3*	c.1049C>T	p.Thr350Met	0	192/0	232	2.80
S1b (OM)	chr1:247,424,492	*NLRP3*	c.1049C>T	p.Thr350Met	7.22	97/7	153	6.90
S2 (PB)	chr1:247,424,357	*NLRP3*	c.914A>C	p.Asp305Ala	36.26	171/62	274	34.80
S3a (PB)	chr16:50,710,912	*NOD2*	c.1001G>A	p.Arg334Gln	10.13	592/60	220	9.40
S3b (OM)	chr16:50,710,912	*NOD2*	c.1001G>A	p.Arg334Gln	5.46	1171/64	349	4.90
S4a (PB)	chr16:50,710,912	*NOD2*	c.1001G>A	p.Arg334Gln	46.44	618/287	231	-
S4 (OM)	chr16:50,710,912	*NOD2*	c.1001G>A	p.Arg334Gln	5.21	576/30	179	8.50
S5a (PB)	chr1:247,425,355	*NLRP3*	c.1912C>G	p.Gln638Glu	19.67	422/83	318	18.40
S5b (OM)	chr1:247,425,355	*NLRP3*	c.1912C>G	p.Gln638Glu	8.72	390/34	274	6.00
S6 (PB)	chr1:247,424,367	*NLRP3*	c.924A>T	p.Gln308His	8.57	175/15	308	5.10
S7a (PB)	chrX:71,109,309	*IL2RG*	c.676C>T	p.Arg226Cys	18.75	192/36	247	17.80
S7b (UR)	chrX:71,109,309	*IL2RG*	c.676C>T	p.Arg226Cys	11.24	169/19	213	8.30
S8 (PB)	chr1:247,424,356	*NLRP3*	c.913G>A	p.Asp305Asn	8.00	125/10	234	7.20
S9 (PB)	chr16:50,710,912	*NOD2*	c.1001G>A	p.Arg334Gln	2.12	1038/22	312	2.70
S10 (PB)	chr14:35,007,365	*SRP54*	c.338G>T	p.Gly113Val	2.34	128/3	146	2.30
S11 (PB)	chr19:855,967	*ELANE*	c.607G>C	p.Gly203Arg	9.10	99/9	219	16.20

VAFs from ADS were extracted from a previous publication (Mensa-Vilaró et al. 2018). DP=total depth; VD=variant depth.

Table 2. Sequential variant filtering process for each sample. The last step where the causal somatic variant is retained is shown in bold.

Filtering	S1a	S1b	S2	S3a	S3b	S4	S5a	S5b	S6	S7a	S7b	S8	S9	S10	S11
On target (VarDict – VarScan2)	298,250 173,072	363,209 101,637	239,526 266,542	286,381 200,507	241,135 453,494	382,519 119,025	231,469 263,604	312,808 245,467	276,040 273,296	293,494 176,720	315,825 171,718	274,664 317,568	191,603 720,791	223,731 261,383	302,540 172,629
Intersection	48,715	44,246	49,871	51,066	70,617	49,399	52,849	61,889	62,644	53,951	51,705	58,201	64,366	50,684	53,268
6pb indels	48,187	43,598	49,246	50,477	69,885	48,451	52,286	61,126	61,954	53,382	50,880	57,646	63,757	50,157	52,621
1000G mask	35,864	31,825	37,231	37,437	55,243	35,592	39,619	47,635	48,200	40,621	38,732	44,829	50,966	37,963	39,983
DP > 50	34,091	27,692	36,459	35,706	54,052	31,989	38,734	45,779	46,668	37,906	35,927	42,349	49,958	33,123	36,816
QUAL > 25	33,771	27,346	35,991	35,272	53,184	31,542	38,295	45,035	45,851	37,388	35,353	41,439	48,835	32,584	36,282
LoF & missense	18,476	15,154	19,998	18,929	31,534	18,100	21,148	26,962	27,596	22,103	20,585	24,283	28,888	18,518	21,472
gnomAD < 0.001	12,135	10,119	13,562	11,980	24,001	12,871	14,427	20,553	21,178	16,281	14,781	17,977	21,910	12,711	15,634
CADD > 15	9,035	7,904	10,085	8,864	19,044	9,994	10,828	16,271	**16,808**	12,662	11,077	13,887	17,086	9,547	12,155
GERP > 2	7,787	6,771	8,703	7,604	16,486	8,582	9,374	13,979	14,498	10,953	9,528	11,976	14,633	8,161	10,409
VD ≥ 3	6,977	6,086	7,446	6,528	14,473	7,720	8,560	12,509	13,231	9,764	8,181	9,719	11,024	5,162	8,991
Candidate genes	174	**177**	**174**	**172**	**319**	**219**	**187**	**276**	275	**226**	**255**	**263**	**243**	**144**	**229**

Table 3. VAF of the 20 somatic candidate variants studied in S5 patient. In grey, values below the sequencing error threshold.

Coordinate (hg38)	Gene	Type	Whole blood	Whole blood post	Urine	Oral mucosa	Neutrophils	Monocytes	B cells	TCD4	TCD8	Control1	Control2	Validated
chr1:153003501	*SPRR3*	Missense	7.7216	3.7419	7.3138	1.6876	7.165	9.5398	0.1088	0.1049	0.0976	0.0832	0.1081	YES
chr1:247425355	*NLRP3*	Missense	24.6228	12.4922	25.3693	7.0918	24.3422	33.993	0.0825	0.0794	0.0787	0.0437	0.0571	YES
chr2:24300108	*ITSN2*	Missense	4.7359	3.4122	5.0154	1.8296	5.6246	2.2179	0.3675	0.1288	0.1415	0.1035	0.1628	YES
chr2:209888127	*UNC80*	Synonimous	26.5932	13.452	28.2002	6.9989	26.8757	34.7143	0.2961	0.2798	0.2896	0.2373	0.2993	YES
chr2:219251622	*TUBA4A*	Synonimous	4.5508	3.2531	4.6625	1.3055	5.3645	1.9301	0.2923	0.1629	0.1553	0.0815	0.1322	YES
chr3:52913506	*SFMBT1*	Missense	1.4477	1.2013	1.3699	0.4748	1.7193	1.453	2.5658	0.1158	0.1569	0.062	0.068	YES
chr4:143695587	*FREM3*	Missense	24.3856	12.176	24.6201	6.1085	23.192	31.5614	0.167	0.1334	0.1238	0.0887	0.1295	YES
chr4:165059454	*TRIM60* *TMEM192*	Intergenic	0.073	0.0596	0.0601	0.0673	0.0659	0.0515	0.0646	0.0892	0.0532	0.0676	0.0533	NO
chr6:36270463	*PNPLA1*	Missense	4.8759	4.0553	5.5917	1.7713	6.4759	2.6609	0.3096	0.0835	0.0768	0.0657	0.0772	YES
chr6:52082518	*PKHD1*	Missense	26.3027	12.7771	27.3696	6.8415	26.0161	35.482	0.1139	0.1225	0.0916	0.0923	0.111	YES
chr6:151349029	*AKAP12*	Missense	1.496	1.2648	1.6739	0.4203	1.8367	1.2469	2.7175	0.0769	0.1867	0.0872	0.0911	YES
chr7:157614060	*PTPRN2*	Intronic	47.7712	48.8889	46.9676	46.2124	49.503	48.0925	48.1535	43.1712	46.6121	0.0656	0.0674	NO
chr9:91410553	*NFIL3*	Missense	0.1427	0.1298	0.127	0.1367	0.109	0.0848	0.1351	0.118	0.1174	0.129	0.1463	NO
chr11:111853480	*ALG9*	Synonimous	4.2723	3.1181	4.6411	1.3842	5.2542	2.2404	0.2136	0.2158	0.229	0.1434	0.2365	YES
chr12:128705237	*TMEM132C*	Missense	2.4998	1.2072	2.4715	0.5706	2.2107	3.4175	0.1087	0.0703	0.0635	0.081	0.0814	YES
chr13:24912928	*CENPJ*	Missense	1.5069	1.1153	1.8446	0.6136	1.8944	0.9268	1.2501	0.1418	0.2347	0.0655	0.0754	YES
chr17:50840691	*WFIKKN2*	Missense	4.1908	2.7357	4.4201	1.3373	4.4152	1.9862	0.1779	0.1208	0.1042	0.0674	0.0797	YES
chr19:16529871	*CHERP*	Synonimous	4.6041	2.2909	4.4651	1.1656	4.6677	5.4904	0.1021	0.1105	0.0973	0.0697	0.0885	YES
chr20:13915139	*SEL1L2*	Intronic	24.5724	11.9959	24.142	6.1982	23.3935	33.2716	0.0564	0.0719	0.0733	0.0562	0.0869	YES
chrX:71537899	*OGT*	Missense	49.978	25.1551	48.1124	12.4303	48.9245	32.2739	0.2688	0.275	0.2301	0.1665	0.2497	YES

Chapter 2

Two fatal cases of deficiency of Interleukin-1 Receptor Antagonist due to novel *IL1RN* variants: a long journey until its definitive diagnosis

Elena Urbaneja MD PhD[1,*] , Manuel Solís-Moruno BsC[2,*], Anna Mensa-Vilaró PhD[3], Iñaki Ortiz de Landazuri-Pascal BsC[3], Susana Plaza AS[3], Virginia Fabregat AS[3], Núria Bonet AS[2], Jordi Yagüe MD PhD[3,4,5], Ferran Casals PhD[2], Juan I. Arostegui MD PhD[3,4,5]

[*] These authors have contributed equally to this work.

Work in progress. Manuscript in preparation.

Abstract

Introduction

The deficiency of IL-1 receptor antagonist (DIRA) is characterized by early-onset severe inflammation mainly affecting skin and bone, and is a consequence of biallelic *loss-of-function* *IL1RN* mutations.

Objective

To elucidate the cause of a lethal disease observed in two siblings with features compatible with DIRA.

Material and methods

Patients' data were collected from their medical charts. Genetic studies were performed in patients' alive relatives using next generation sequencing (NGS) methods. Relative *IL1RN* expression and mRNA sequencing will be performed to characterize the transcriptional consequences of the detected variants.

Results

NGS studies detected two novel heterozygous *IL1RN* variants in each one of the patients' parents. The father's variant was located at the donor splicing site of intron 3 (c.318+2T>G), and subsequent studies will be performed to reveal a predicted decrease of gene transcription compared with healthy controls, which would strongly suggest the production of a non-functional *IL1RN* allele. By contrast, the mother's variant was a large intragenic deletion (~2500bp) encompassing from intron 1 until exon 3, and it expected to generate a truncated protein. *IL1RN* genotypes of all patients' alive relatives were compatible with a recessive inheritance pattern for the disease.

Conclusions

Two novel *IL1RN* variants predicted to generate truncated proteins were identified in this family. The non-availability of patients' samples was a limitation to establish their definitive DIRA diagnosis. However, their clinical features, the familial recurrence of the disease and the genetic evidences here shown strongly suggest that they suffered from a lethal form of DIRA, most probably as a consequence of compound heterozygous *IL1RN* genotypes.

Introduction

Interleukin-1 (IL-1) was first described in early 70s as a soluble factor related to lymphocyte activation, emergency hematopoiesis and fever (Gery, Gershon, and Waksman 1972). At present, it represents the prototypic cytokine driving local and systemic inflammation. There are two different IL-1 proteins named IL-1α and IL-1β, which are encoded by two genes (*IL1A* and *IL1B*) located at chromosome 2q14. Despite the marked differences they have related to their amino acid sequences, tissue expression and post-translational modifications, their active forms bind to the IL-1 receptor type I (IL-1RI), recruit the co-receptor IL-1R accessory protein (IL-1RAcp) and transduce an intracellular signal that triggers the production of inflammatory mediators. A sustained activity of this IL-1-mediated inflammatory cascade may have deleterious consequences for the normal homeostasis. Consequently, there are different mechanisms that maintain under control its overall activity. The IL-1 receptor antagonist (IL-1Ra), a member of the family of IL-1 cytokines, represents one of the subtlest mechanisms of control of this pathway (Seckinger et al. 1987). This protein binds to the IL-1RI, but it does not recruit the IL-1RAcP, and does not transduce a pro-inflammatory signal. As an overall consequence, IL-1Ra competitively antagonizes the pro-inflammatory action of the binding of IL-1α and IL-1β to the IL-1RI (Dayer, Oliviero, and Punzi 2017).

IL-1Ra is encoded by *IL1RN*, a gene located in the IL-1 cluster of genes at chromosome 2q14. Biallelic *loss-of-function IL1RN* mutations have been described in both humans and mice causing the recessively inherited deficiency of IL-1Ra (DIRA) (Aksentijevich et al. 2009). In humans, DIRA is a disease that typically starts early in life, and progresses chronically, with pustular skin rash, sterile bone inflammation and increased levels of acute phase reactants as its main features. The elucidation of its molecular basis leads to treating these patients successfully with anakinra, the recombinant form of human IL-1Ra.

As occurs with all inherited diseases, DIRA has probably affected a small number of patients in the past. However, since the description of its molecular basis in 2009, around 20 cases have been reported, thus expanding the clinical and genetic diversity of the disease. In the present study, we describe a fatal and devastating inflammatory disease observed in two siblings from a non-consanguineous couple of Spanish ancestry, with clinical features consistent with DIRA. Genetic investigations were performed three decades after the patients' death using samples from their parents and siblings, which identified two rare and previously unrecognized heterozygous *IL1RN* variants. They are a single nucleotide variant at a donor splicing site and a ~2600 bp genomic deletion. Additional experiments will show the deleterious consequences of both variants at mRNA expression level, suggesting they generate truncated proteins. This evidence strongly suggests that the patients here described suffered from DIRA and died due to the natural course of the disease, which most probably occurred as a consequence of compound heterozygous genotypes for the rare variants at the *IL1RN* locus.

Material and methods

Patients

We identified a non-consanguineous couple of Spanish ancestry with four children, two affected siblings who died in the 80s of an undiagnosed and severe inflammatory disease, and two healthy and alive siblings (Figure 1A). Clinical data and results of analytical tests were collected after review of medical charts. Written informed consents for molecular analyses were obtained from each enrolled individual. The Ethical Review Boards Hospital Clínico Universitario de Valladolid and Hospital Clínic, Barcelona, all in Spain, approved the study. All investigations were performed in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments.

Genomic Analyses

Genomic DNA samples from patients' healthy parents and siblings were prepared from unfractionated peripheral blood using a QIAmp DNA Blood Mini Kit (QIAgen, Germany).

A first genetic study by targeted gene panel sequencing including all genes associated with monogenic autoinflammatory diseases was performed for short variant discovery, i. e., SNVs and indels.

For Sanger sequencing, all exons and intronic boundaries of *IL1RN* gene (RefSeq NM_173842.3) were PCR-amplified, purified with Illustra ExoStar 1-Step kit (GE Healthcare, USA), bidirectionally fluorescence sequenced using ABI BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, USA) and run on an automated ABI 3730XL DNA analyzer (Applied Biosystems, USA). Reads analyses were performed with the SeqPilot software (JSI Medical Systems, Germany). To determine the specific deletion breakpoint, a PCR amplicon using primers located at each boundary of the deletion was designed and subsequently sequenced as previously described.

Whole Genome Sequencing

We sequenced the whole genome of the patient's mother in order to expand the analysis of SNVs and indels and to explore its structural variant landscape. Illumina 2 × 150 paired-end cycles runs to a mean coverage ~37X, using the TruSeq Nano DNA (350) library kit, was performed in Macrogen facilities.

We mapped raw reads with BWA-MEM (Li 2013) algorithm (v. 0.7.16a-r1181) to the human reference genome GRCh38/hg38 (UCSC), marked duplicated reads with MarkDuplicates from picard tools (v. 2.18.6), performed indel realignment with IndelRealigner from GATK (McKenna et al. 2010) (v. 3.7-0-gcfedb67) and base quality score recalibration with BQSR tool, also from GATK. We called SNVs and indels with GATK's HaplotypeCaller. We finally annotated the genetic variants using SnpEff (Cingolani, Platts, et al. 2012) (v. 4.3t) and SnpSift (Cingolani, Patel, et al. 2012) (v. 4.3t).

For the structural variant calling we used CNVnator (Abyzov et al. 2011) (v. 0.3.3) with bin sizes of 100 and 50 bp.

Results

IL1RN **genomic analyses**

On the basis of patients' clinical data, laboratory results and familial pedigree, we hypothesized for the presence of a severe, recessively-inherited inflammatory disease in this family, with DIRA as the most probable candidate disease. As biological samples from patients were not available for this study, we obtained and genetically analyzed samples from their healthy parents and siblings. A first genetic study using a targeted gene panel sequencing for monogenic autoinflammatory diseases revealed in the patients' father and brother a heterozygous T>G transversion at position +2 of the donor acceptor splicing site of intron 3 of the *IL1RN* gene (Figure 1B). The c.318+2T>G *IL1RN* variant has not been previously reported, is absent in all available databases (1000 Genomes Project, NHBLI-ESP, gnomAD, Collaborative Spanish Variant Server) and bioinformatics analyses predicted to impair the normal splicing of *IL1RN* mRNA (Table 1). Altogether, these evidences strongly suggest that it may be a pathogenic *IL1RN* variant and it will be subject of subsequent analyses. By contrast, no variants were detected in patients' mother and sister at the *IL1RN* locus.

As DIRA diagnosis may fit with the clinical features observed in the patients, and considering the intrinsic limitations of the sequencing methods previously employed, we performed additional genetic analyses to identify the second hit at the *IL1RN* locus. Whole genome sequencing was performed in the sample from patients' mother, which revealed a potential heterozygous, intragenic genomic deletion encompassing ~2500 bp (Figure 1C). In order to characterize the ends of this large deletion, a specific PCR amplification with primers at each side of the breakpoint was designed (Figure 1D). The sequence of this amplicon revealed a 2592 bp deletion at the *IL1RN* locus in the patients' mother (Figure 1E), which encompasses from middle of intron 1 until the middle of exon 3 (Supplemental Figure S1). This deletion has not been previously reported in patients with DIRA neither have been registered in all available databases. Prediction analyses suggest that the *IL1RN* allele containing this genomic deletion produces a transcript smaller than that encoded by the normal allele, and supporting for a non-functional protein. Intrafamilial genotypes for the *IL1RN* deletion were obtained in all patients' relatives by a specific PCR amplification, revealing that it was exclusively present in the patients' mother as a single heterozygous genotype (Figure 1F).

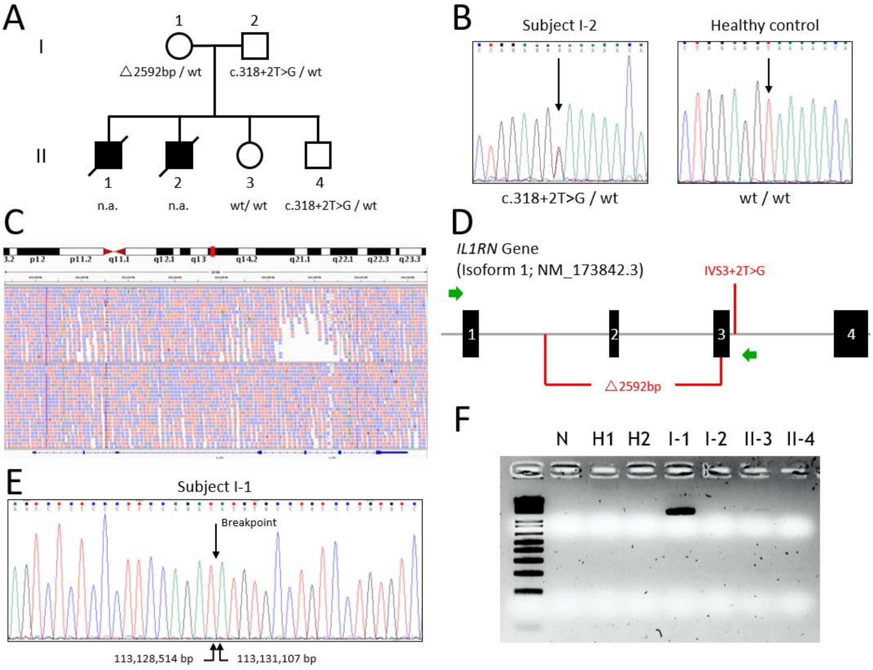


Figure 1. Genomic *IL1RN* Variants Detected in Enrolled Individuals. Panel A. Pedigree of family. Black filled symbols represent affected subjects, open symbols unaffected subjects, squares male subjects, circles female subjects, and slashed deceased subjects. *IL1RN* genotypes are shown below each analyzed subject. n.a., not analyzed; wt, wild-type. **Panel B.** Sense Sanger chromatograms from subject I-2 carrying the heterozygous genotype for the c.318+2T>G *IL1RN* variant (left box) and from a healthy subject (right box). The arrows indicate the position where the nucleotide variant is located. **Panel C.** Mapped reads from whole genome sequencing performed in subject I-1 showing the loss of coverage at *IL1RN* locus, suggesting the presence of an intragenic, heterozygous genomic deletion. In the bottom part of the panel, an unrelated individual sequenced with the same protocol is shown for comparison. Reads were visualized using Integrative Genomics Viewer. **Panel D.** Genomic organization of isoform 1 of *IL1RN* gene. Variants identified in the enrolled family are shown in red. Green arrows represent the forward and reverse primers designed to generate a PCR amplicon to specifically identify the presence of the genomic deletion. **Panel E.** Sense Sanger chromatogram showing the breakpoint and boundaries of the genomic deletion at the *IL1RN* locus identified in subject I-1. The arrows indicate the nucleotides located at each side of the breakpoint site, and below are shown the respective genomic coordinates according to GRCh38. **Panel F.** Agarose gel electrophoresis of PCR products generated with the use of primers designed for the genomic deletion. N, negative control; H1 and H2, healthy control subjects. Black arrows indicate the specific bands of PCR amplicons containing the *IL1RN* deletion (top) and a positive control of PCR reaction (bottom).

Discussion

Our results suggest that both newly described variants in the *IL1RN* gene, c.318+2T>G and the 2592 bp deletion, cause DIRA when occurring together. Our study presents two main limitations: we do not have available DNA from any of the patients and we have not performed functional analysis yet. For the first case, we are not able to confirm whether both variants were actually found together. In the second case, we cannot assure if any of the mutations really disrupt the protein, although we plan to perform analyses at the mRNA level to confirm that its production is affected.

Another deletion encompassing the *IL1RN* gene plus five other genes of the IL-1 family has been reported (Reddy et al. 2009). This homozygous 175 kb deletion at chromosome 2q13 was described by the authors as the causative agent of a disease different from the neonatal-onset multisystem inflammatory disorder (NOMID). The symptomatology of this patient was effectively controlled by anakinra. This work was published at the same time than the one by Aksentijevich et al. (Aksentijevich et al. 2009), in which the term DIRA was coined.

Acknowledgments

This work has been partially funded by CERCA Programme/Generalitat de Catalunya (JI.A.). Funded by RTI2018-096824-B-C21 grant from the Spanish Ministry of Science, Innovation and Universities co-financed by European Regional Development Fund (JI.A.), by grant SAF2015-68472-C2-2-R from the Ministerio de Economía y Competitividad (Spain) and by Direcció General de Recerca, Generalitat de Catalunya (2017SGR-702) (F.C.). M.S.-M. is supported by the Ministerio de Economía y Competitividad, Spain (Maria de Maetzu grant MDM-2014-0370-16-3).

Bibliography

Abyzov, Alexej, Alexander E. Urban, Michael Snyder, and Mark Gerstein. 2011. "CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing." *Genome Research* 21 (6): 974–84. https://doi.org/10.1101/gr.114876.110.

Aksentijevich, Ivona, Seth L. Masters, Polly J. Ferguson, Paul Dancey, Joost Frenkel, Annet van Royen-Kerkhoff, Ron Laxer, et al. 2009. "An Autoinflammatory Disease with Deficiency of the Interleukin-1–Receptor Antagonist." *New England Journal of Medicine* 360 (23): 2426–37. https://doi.org/10.1056/nejmoa0807865.

Cingolani, Pablo, Viral M. Patel, Melissa Coon, Tung Nguyen, Susan J. Land, Douglas M. Ruden, and Xiangyi Lu. 2012. "Using Drosophila Melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift." *Frontiers in Genetics* 3 (MAR): 1–9. https://doi.org/10.3389/fgene.2012.00035.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff." *Fly* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

Dayer, Jean Michel, Francesca Oliviero, and Leonardo Punzi. 2017. "A Brief History of IL-1 and IL-1 Ra in Rheumatology." *Frontiers in Pharmacology* 8 (MAY): 1–8. https://doi.org/10.3389/fphar.2017.00293.

Gery, Igal, Richard K. Gershon, and Byron H. Waksman. 1972. "Potentiation of the T-Lvmphocyte Response to Mitogens: I. The Responding Cell." *Journal of Experimental Medicine* 136 (1): 128–42. https://doi.org/10.1084/jem.136.1.128.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *ArXiv:1303.3997 [q-Bio.GN]*, March. http://arxiv.org/abs/1303.3997.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Reddy, Sreelatha, Shuang Jia, Rhonda Geoffrey, Rachel Lorier, Mariko Suchi, Ulrich Broeckel, Martin J. Hessner, and James Verbsky. 2009. "An Autoinflammatory Disease Due to Homozygous Deletion of the IL1RN Locus." *New England Journal of Medicine* 360 (23): 2438–44. https://doi.org/10.1056/NEJMoa0809568.

Seckinger, P, J W Lowenthal, K Williamson, J M Dayer, and H R MacDonald. 1987. "A Urine Inhibitor of Interleukin 1 Activity That Blocks Ligand Binding." *Journal of Immunology* 139 (5): 1546–49. http://www.ncbi.nlm.nih.gov/pubmed/2957429.

Table 1. Details of the intronic variant detected in the *IL1RN* gene.

Chromosome Position[1]	**Gene[2]**	**Intron**	**cDNA alteration**	**Amino acid alteration**	**Population Genetics**				**Bioinformatics**			**Variant Classification[3]**
					1000 GP	NHLBI ESP	gnomAD	CSVS	Human Splicing Finder	GERP score	CADD PHRED	
chr2:113131159	*IL1RN*	3	c.318+2T>G	-	0	0	0	0		6.41	33	Pathogenic

[1]Referred to GRCh38. [2]RefSeq: NM_173842.3. [3]Classification of pathogenicity of gene variants performed on the basis of standards and guidelines proposed in the consensus recommendations of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Abbreviations: chr, chromosome; 1000 GP, 1000 Genomes Project Phase 3; NHLBI-ESP, National Heart, Lung and Blood Institute-Exome Sequencing Project; gnomAD, Genome Aggregation Database; CSVS, Collaborative Spanish Variant Server.

Chapter 3

Flow sorting enrichment and Nanopore sequencing of chromosome 1 from a Chinese individual

Lukas F. K. Kuderna[1†*], Manuel Solís-Moruno[1,2†], Laura Batlle-Masó[1,2†], Eva Julià[3,4†], Esther Lizano[1†], Roger Anglada[2], Erika Ramírez[4], Alex Bote[4], Marc Tormo[2,5], Tomàs Marquès-Bonet[1,6,7,8,9], Òscar Fornas[4,10†*] and Ferran Casals[2†*]

†These authors have contributed equally to this work.

Frontiers in Genetics 10 (January): 1–6.

doi: 10.3389/fgene.2019.01315

Abstract

Sorting of individual chromosomes by Flow Cytometry (flow-sorting) is an enrichment method to potentially simplify genome assembly by isolating chromosomes from the context of the genome. We have recently developed a workflow to sequence native, unamplified DNA and applied it to the smallest human chromosome, the Y chromosome. Here, we modify and improve upon that workflow to increase DNA recovery from chromosome sorting as well as sequencing yield. We apply it to sequence and assemble the largest human chromosome –chromosome 1– of a Chinese individual using a single Oxford Nanopore MinION flow cell. We generate a selective and highly continuous assembly whose continuity reaches into the order of magnitude of the human reference GRCh38. We then use this assembly to call candidate structural variants against the reference and find 685 putative novel SV candidates. We propose this workflow as a potential solution to assemble structurally complex chromosomes, or the study of very large plant or animal genomes that might challenge traditional assembly strategies.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 09 January 2020
doi: 10.3389/fgene.2019.01315

Flow Sorting Enrichment and Nanopore Sequencing of Chromosome 1 From a Chinese Individual

Lukas F. K. Kuderna[1†*], *Manuel Solís-Moruno*[1,2†], *Laura Batlle-Masó*[1,2†], *Eva Julià*[3,4†], *Esther Lizano*[1†], *Roger Anglada*[2], *Erika Ramírez*[4], *Alex Bote*[4], *Marc Tormo*[2,5], *Tomàs Marquès-Bonet*[1,6,7,8,9], *Òscar Fornas*[4,10†*] and *Ferran Casals*[2†*]

[1] Institut de Biologia Evolutiva, CSIC-Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona (PRBB)", Barcelona, Spain, [2] Genomics Core Facility, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona (PRBB), Barcelona, Spain, [3] Serveis Cientifico-Tècnics, Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Barcelona, Spain, [4] Flow Cytometry Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology (BIST), Barcelona, Spain, [5] Scientific IT Core Facility, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona (PRBB), Barcelona, Spain, [6] CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain, [7] Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), Barcelona, Spain, [8] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, [9] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain, [10] Flow Cytometry Unit, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), Parc de Recerca Biomèdica de Barcelona (PRBB), Barcelona, Spain

OPEN ACCESS

Edited by:
Youri I. Pavlov,
University of Nebraska Medical Center,
United States

Reviewed by:
Satomi Mitsuhashi,
Yokohama City University,
Japan
Yael Michaeli,
Tel Aviv University,
Israel

***Correspondence:**
Lukas F. K. Kuderna
lukas.kuderna@upf.edu
Òscar Fornas
oscar.fornas@upf.edu
Ferran Casals
ferran.casals@upf.edu

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: 18 September 2019
Accepted: 02 December 2019
Published: 09 January 2020

Citation:
Kuderna LFK, Solís-Moruno M, Batlle-Masó L, Julià E, Lizano E, Anglada R, Ramírez E, Bote A, Tormo M, Marquès-Bonet T, Fornas O and Casals F (2020) Flow Sorting Enrichment and Nanopore Sequencing of Chromosome 1 From a Chinese Individual. Front. Genet. 10:1315. doi: 10.3389/fgene.2019.01315

Sorting of individual chromosomes by Flow Cytometry (flow-sorting) is an enrichment method to potentially simplify genome assembly by isolating chromosomes from the context of the genome. We have recently developed a workflow to sequence native, unamplified DNA and applied it to the smallest human chromosome, the Y chromosome. Here, we modify improve upon that workflow to increase DNA recovery from chromosome sorting as well as sequencing yield. We apply it to sequence and assemble the largest human chromosome - chromosome 1 - of a Chinese individual using a single Oxford Nanopore MinION flow cell. We generate a selective and highly continuous assembly whose continuity reaches into the order of magnitude of the human reference GRCh38. We then use this assembly to call candidate structural variants against the reference and find 685 putative novel SV candidates. We propose this workflow as a potential solution to assemble structurally complex chromosomes, or the study of very large plant or animal genomes that might challenge traditional assembly strategies.

Keywords: chromosome enrichment, nanopore sequencing, chromosome sequencing, chromosome sorting, flow karyotyping, structural variation, genome assembly

INTRODUCTION

Structural genetic variation is abundant and has important functional impact (Conrad et al., 2010). A human genome has been estimated to harbor more than 2,000 structural variants (SV), which are typically defined as variants that affect at least 50 bp (Mills et al., 2011). They include balanced (inversions, translocations) and unbalanced forms (insertions, deletions, duplications) (Mills et al., 2011). The functional impacts are mainly produced by altering the number of copies of coding sequences, and thus the gene expression levels, or disrupting coding or regulatory regions with a potential effect not only in the closer genes but also extending to hundreds of kilobases away (Weischenfeldt et al., 2013). SVs have been associated both to Mendelian and common disorders, although it is difficult to exactly define the phenotypic impact due to the presence of many functional regions in the same variant, as well as variable expressivity and penetrance even across family (Weischenfeldt et al., 2013). Nevertheless, and despite of its high abundance and potential impact, the study of SVs has been less accelerated in comparison to single nucleotide variants and short insertions and deletions (indels). It has been mainly limited by the short reads generated by massive parallel sequencing technologies and the relatively low coverage in large sequencing efforts (e.g., 1,000 Genomes Project) (Huddleston and Eichler, 2016). Also, determining the exact position and mechanism of origin of SVs is not straightforward often due to the presence of terminal repetitive sequences and recurrence, and can be especially challenging in complex structural variants with more than two breakpoints and overlapping or nested rearrangements (Collins et al., 2017; Stephens et al., 2018). All this makes difficult the generation of systematic catalogues of SVs and the estimation of allelic population frequencies.

The recently emerging possibility to obtain reads of up to of several Megabases in length on the Oxford Nanopore platform represents an important advance for the study of structural variants and genome assembly, as it greatly simplifies them (Giordano et al., 2017; Payne et al., 2019). These sequencing technologies can be combined with enrichment strategies, from capture by hybridization to Cas9 based methodologies, to restrict the analysis to specific regions also increasing the sequencing yield. Chromosome isolation is an alternative enrichment strategy which will better maintain molecular integrity with the potential of generating longer sequence reads (Jiang et al., 2015; Kozarewa et al., 2015; Gabrieli et al., 2018).

We recently developed a workflow to isolate and sequence native flow-sorted human Y chromosomes on an Oxford Nanopore MinION device (Kuderna et al., 2019). We sought to apply this method to other chromosomes to generate a population specific long read assembly, namely for a chromosome 1 of a Chinese individual. We show the generalizability and improve the protocol in terms of DNA recovery and sequencing yield.

METHOD

Chromosome Isolation and Sequencing

Chromosome preparation, staining, sorting, DNA purification, concentration and sequencing were performed as previously described in (Kuderna et al., 2019) with some modifications (see supplementary methods). Briefly, chromosomes were prepared from a lymphoblastoid cell line derived from a Chinese individual (Coriell, cat. no. HG00542) by using a polyamine isolation method. Modifications: hypotonic solution was incubated at 37°C for 20 min and polyamine isolation buffer was incubated on ice for 30 min. Additionally, potassium citrate was replaced by sodium citrate and sodium sulfite to a final concentration of 10 and 25 mM respectively and incubated at least 2 hours to enhance peak resolution in the flow karyotype. Purification and concentration were carried out as previously described with the exception that after SPRI bead purification DNA was eluted in 20 µl of Low TE buffer. Libraries for sequencing were prepared from the purified DNA following the protocol of the Ligation Sequencing Kit SQK-LSK 109 (Oxford Nanopore Technologies). A 48 hours MinION run was performed in a FLO-MIN106 flow cell.

Assembly, Error Correction, and SV Calling

We called bases from the raw fast5 signal data using Guppy (v. 2.2.2) with the following command line:

```
  guppy_basecaller -i $input -s $output –
flowcell FLO-MIN106 –kit SQK-LSK109 -t 4
–disable_pings
```

We mapped the basecalled reads onto GRCh38 using Minimap2 (Li, 2018) with the ont preset. We sorted the mappings and converted them to bam using samtools (Li et al., 2009):

```
  minimap2 -x map-ont -t8 -a hg38.fa
basecalled_reads.fq| samtools sort -@8 -O BAM
- -o mapped_reads.bam
```

We unsuccessfully tried to assemble the raw reads into contigs using canu (v. 1.8) (Koren et al., 2017) with default parameters assuming a "genome" size of 250 Mb. This command used more than 15 Tb of disk space and did not finish to yield a successful assembly on our systems. To overcome the issue, we extracted mappings on chromosome 1 and assembled only those:

```
  canu -p HG00542-chr1 -d HG00542-chr1
genomeSize = 250m -nanopore-raw
basecalled_reads.chr1_mappings.fq
```

We corrected errors in the resulting contigs with Nanopolish (v. 0.11.0) (Simpson et al., 2017). To this end, we remapped the raw reads to the assembly as shown above. We then went on to create a read db with nanopolish, and split the assembly into chunks of 500 Kb with nanopolish_makerange.py and called the variants of each chunk with nanopolish variants

```
  nanopolish_makerange.py –segment-length
500000 –overlap-length 1000 HG00542- HG00542-
chr1.contigs.fasta | xargs -i echo nanopolish
variants –ploidy 2 –consensus -o
{}.consensus.round1.vcf -w {} -r
basecalled_reads.fq -b HG00542- HG00542-
chr1.contigs.self-mappings.bam -g HG00542-
chr1.contigs.fasta | sh
```

We then incorporated the corrections into the assembly:

```
  nanopolish vcf2fasta -g HG00542-
chr1.contigs.fasta *vcf.
```

We aligned the resulted polished assembly to GRCh38 chr1 with MUMmer (Kurtz et al., 2004)

```
nucmer -maxmatch -l 100 -c 500 hg38.
chr1.toplevel.fa. HG00542-chr1.contigs.
polished.fasta -prefix HG00542.
polished.r1.vs.hg38_chr1
```

We fed the resulting alignments to Assemblytics to obtain candidates for SV

```
Assemblytics HG00542.polished.r1.vs.hg38_
chr1.delta HG00542.polished.r1.vs.
hg38_chr1.10kanchor.50kmax 10000
bin/Assemblytics/
```

We generated an additional callset with Sniffles (v. 1.0.8) (default parameters), using the previously mapped reads from minimap2. For downstream analysis we only retained calls annotated as "precise" by Sniffles:

```
sniffles -m mapped_reads.bam -v
sniffles_callset.vcf.
```

We filtered all calls that fall within 2 Mb of distance to the centromere or telomeres.

Comparative Repeat Annotation

We ran repeatmasker (v. 4.0.7) with the same parameters on both our assembly and GRCh38 to create comparable annotations:

```
RepeatMasker -e ncbi -pa 8 -s -species human
-no_is -noisy -dir. -a -gff -u assembly.fa
```

We calculated the divergence of each repeat to its consensus using the "calcDivergenceFromAlign.pl" utility included in the RepeatMasker package.

RESULTS AND DISCUSSION

We sorted 10 million individual chromosomes 1 from a lymphoblastoid cell line derived from a Chinese individual (HG00542) to obtain 5 µg of DNA from a total of 205x10^6 cultured cells from six independent experiments (see **Supplementary Figure 1**). Of that, we used 2 million sorted chromosomes theoretically corresponding to 1µg of DNA (Gribble et al., 2009). From this material, we constructed a library using an Oxford Nanopore ligation kit and ran a single MinION flow cell on it. Given limitations in DNA recovery from flow-sorted material we have previously encountered, we have made adjustments to the sorting and purification protocol (see methods and supplementary methods). The higher DNA recovery and higher loading amount on the flow cell yielded between 20 and 131 times more data from a single run than our previous efforts, meaning that a single flow cell is now sufficient to assemble the largest human chromosomes after flow-sorting it. These differences are likely also attributable to improvements in the pore chemistry and base-calling algorithms. After base calling with Guppy, we were left with 2.5 million reads summing to 14.3 Gb of data with a read length N50 of 15.4 Kb. Of them, 10.6 Gb mapped readily to GRCh38, and 5.6 Gb to chromosome 1 (see **Supplementary Figure 2**). The average coverage on chromosome 1 was 28.4 fold, the coverage on the remaining chromosomes ranged from 0.7 fold on chrX to 2.1 fold on chr19 (**Figure 1**). This amounts to an 8-fold enrichment over a random sampling from bases along a diploid male genome (see **Supplementary Table 5**). All other chromosomes are depleted from the data, with depletion ranging from 0.27 fold on chr4 to 0.61 fold on chrY. We find the average depletion on non-target chromosomes to be more efficient in this dataset compared to our previous effort on the Y chromosome (0.42 fold versus 0.61 fold). Nevertheless, we observe the enrichment on the target chromosome to be less efficient compared to the Y chromosome. This fact is likely attributable to the more challenging physical separation of chromosome 1 in a human flow karyogram, as the chromosome clusters are not as well defined as e.g. the one of chrY (**Supplementary Figure 1**).

We assembled the raw data using Canu (Koren et al., 2017). To this end, we removed reads that do not map to GRCh38 chr1 to ease the computational load of the assembly (see Method). While this might confound the assembly in regions of large insertions or translocations, it significantly eases computational burden. We polished remaining single base substitution and indel errors in the resulting assembly with Nanopolish (Simpson et al., 2017). The final assembly has a length of 227.8 Mb and consists of 154 contigs with an N50 of 10.5 Mb. We aligned our assembly to GRCh38 chromosome 1, whose total resolved sequence length (i.e. excluding "N" from the assembly) is 230.5 Mb. We find 98.8% of our assembly to cover 97.6% of the reference (**Figure 2**, **Supplementary Table 4**). The boundaries of our contigs are enriched in segmental duplications and satellite repeats in the reference. We observe the highest degree of fragmentation around the centromeric region, which is littered with satellite repeats and segmental duplications, and therefore particularly challenging to assemble. Contigs mapping to these regions also show a drop off in identity to the reference. The centromere on chromosome 1 of GRCh38 is an 18 Mb long heterochromatic expansion flanked by segmental duplications that is still unresolved, as in most other human chromosomes (Jain et al., 2018).

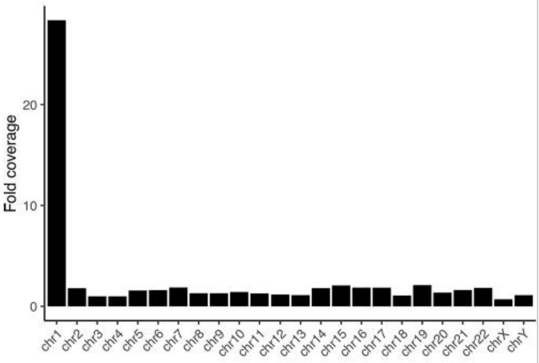


FIGURE 1 | Fold coverage per chromosome. The sequencing is selective for chromosome 1, with all other chromosomes being depleted from the data.

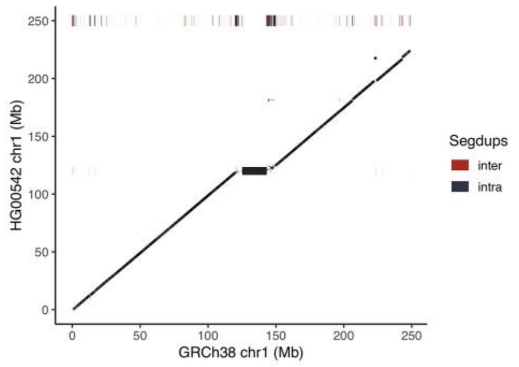


FIGURE 2 | Dot-plot of HG00542 assembly versus GRCh38 chromosome 1. The chromosomes are laid out on the respective axis and a dot denotes aligned sequence between the two assemblies. Bars at the height of 250 Mb on the Y axis show the positions of segmental duplications in GRCh38. The assembly is largely colinear to the reference. The large black block in the center of the dot-plot delimits the 18 Mb centromere of chromosome 1.

To assess repeat resolution, we produced a comparative repeat annotation between our assembly and GRCh38 using repeatmasker (Smit et al.). We find both assemblies to have very similar proportions annotated as repetitive overall and for all given repeat families. We then calculated the divergence of all annotated repetitive elements to their consensus sequence to create "repeat landscapes". We find these landscapes to be highly similar between the two assemblies. We measured repeat resolution in our assembly as the proportion of bases annotated as a given repeat type. We find them to be of comparable quality across all major repeat types, with centromeric & telomeric satellite sequences constituting an exception (**Figure 3**, **Supplementary Figure 3**, **Supplementary Table 1**).

We then used the assembly to generate a call set of candidate structural variants (SVs) against GRCh38. To this end, we aligned our assembly to the reference using MUMmer (Kurtz et al., 2004) and looked for patterns of SVs using Assemblytics (Nattestad and Schatz, 2016). We additionally ran an orthogonal detection approach by mapping the raw reads to GRCh38 and running Sniffles (Sedlazeck et al., 2018). To minimize erroneous calls, we excluded putative SVs within 2 Mb of the centromeric and telomeric regions, as the higher degree of segmental duplications and assembly fragmentation is more likely to yield false positive calls (Audano et al., 2019). By this means, we identified 1,325 SVs with Assemblytics and 940 with Sniffles along chromosome 1. We find 405 of the calls to intersect between the two sets, with 61.4% and 56.9% to be unique to Assemblytics and Sniffles, respectively (see **Supplementary Figures 4–8**). Of the intersecting calls, we find 230 to lie within genic regions, and 8 to affect the coding portions of the gene (**Figure 4** and **Supplementary Tables 2–3, 6–7**).

We sought to assess novel SVs on the one hand, and population frequencies of SVs that might have previously been described in other datasets. To this end, we contrasted our calls against those generated by the 1,000 genomes consortium (Sudmant et al., 2015), which used short-read data to detect SVs with several different algorithms. This study detected 4,653 SVs on chr1 among 2,504 individuals. Unsurprising, given the technological differences between the two datasets, we find comparatively little overlap between the two call sets with 466 SVs that overlap over 40% in either of them. We calculated the frequencies of overlapping SVs in each of the superpopulations of the 1,000 genomes data. After removing variants with an allele count of 2 or less, and multiallelic positions we find these SVs to reach the highest frequencies in east Asian populations (20.5%); South Asian and American populations exhibit similar frequencies (18.8% and 18.4%) followed by European (14.7%) and lastly African (9.8%) populations (see **Figure 5** and **Supplementary Figures 9–12**). We additionally contrasted our calls with more recently generated ones that also used long-read assemblies for detection (Audano et al., 2019). Among 15 individuals included in that study there are 6,646 SVs

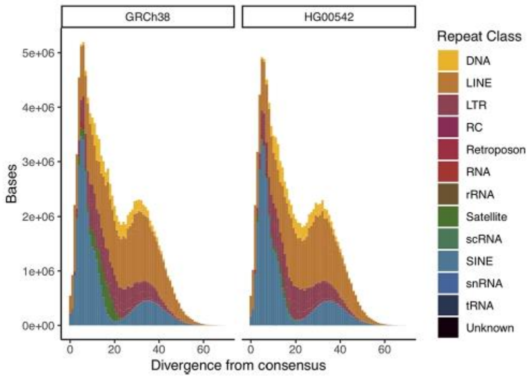


FIGURE 3 | Comparative repeat landscapes of GRCh38 chromosome 1 and HG00542 chromosome 1. We find equal resolution across most repeat classes.

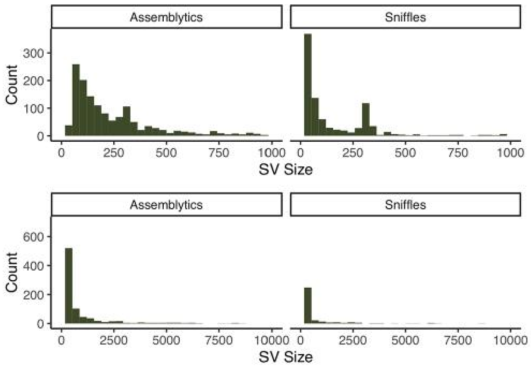


FIGURE 4 | Size distribution of SV calls from both Assemblytics and Sniffles at different resolutions. Both call sets have clear peaks around 300 bp corresponding to Alu-elements.

along chr1. We find 291 SVs from our call set to overlap those. The calls by Audano et al. include a Chinese individual (HX1) and one from Korea (AK1) (Seo et al., 2016; Shi et al., 2016). Among the overlapping SVs, we find 108 deletions (or 44%) and 32 insertions (64%) to also be present in both these individuals. Lastly, we identify 685 novel SV candidate loci that have not previously been described in neither of the above datasets.

In summary, we generated a highly continuous and selective assembly of the largest human chromosome from a Chinese individual from flow-sorted native DNA. We show that increased efficiency in DNA recovery from flow-sorted chromosomes, as well as improvements in nanopore technology, allow for single chromosome assemblies from a single MinION flow cell and that as little as 28-fold coverage is sufficient to yield an assembly with a contig N50 over 10 Mb. As with previous reports, we still find room for improvement in terms of base accuracy. We observe a deletion bias in our data, which we find to be twice as frequent as insertions. However, given the constant development in both pore design and base calling algorithms, these issues are likely to improve in the near future.

It is worth noting that flow-sorting chromosomes only constitutes a viable approach if the species' chromosomes are sufficiently distinct in terms of size and GC content. As an example, human chromosomes 9–12 have size differences of up to 6%. However, with our approach, they are hardly distinguishable by flow karyotyping because of similar GC-content across them. Conversely, human chromosomes 1–2 have a size difference of only 1.6%. Nevertheless, they differ more strongly in GC content, making them clearly distinguishable by flow karyotyping. Addressing this “sortability” of a species' genome is achieved empirically. While assembling mammalian genomes has become routine, there is still a large amount of plants and animals for which traditional whole-genome shotgun assembly methods might be computationally prohibitive given their massive genome sizes. We expect assembling flow-sorted chromosomes to be a viable alternative in these cases.

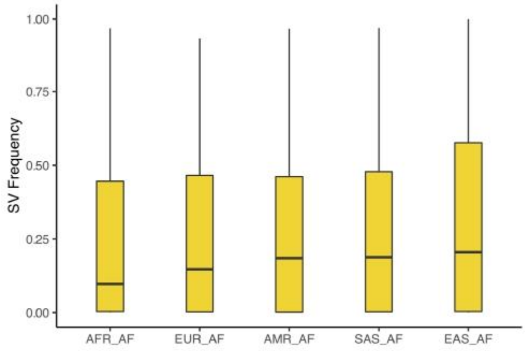


FIGURE 5 | Population frequencies of SV discovered in our assembly that overlap calls by the 1,000 genomes project. We find these SVs to be most frequent in East Asian populations, followed South Asian and American, European, and lastly African populations.

DATA AVAILABILITY STATEMENT

The sequencing data has been deposited at the European Nucleotide Archive (ENA) under the accession PRJEB34445. The assembly can be accessed under GCA_902652775.1.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LK, MS-M, LB-M, TM-B, OF, and FC conceived the project. LK, MS-M, LB-M, EJ, EL, OF, and FC designed the study. LK, MS-M, LB-M, and MT performed the bioinformatic analysis. EJ, EL, RA, ER, and AB performed the experimental analysis. All the authors participated in the analysis of the data. LK, MS-M, LB-M, EJ, OF, and FC wrote the manuscript.

FUNDING

This study was funded by grants RTI2018-096824-B-C22 from the Agencia Estatal de Investigación-Ministerio de Ciencia, Innovación y Universidades (Spain) and FEDER (EU) to OF and FC, SAF2015-68472-C2-2-R from the Ministerio de Economía y Competitividad (Spain) and FEDER (EU) to FC, the Centro de Excelencia Severo Ochoa, and by Direcció General de Recerca, Generalitat de Catalunya (2017SGR-702). TM-B is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social “La Caixa” and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). LK is supported by an FPI fellowship associated with BFU2014-55090-P (MINECO/FEDER, UE) and by an EMBO Short-Term Fellowship STF-8286. LB-M is supported by a Formació de personal Investigador fellowship from Generalitat de Catalunya (2018_FI_B00072). MS-M is supported by the María de Maetzu Programme (MDM-2014-0370-16-3).

ACKNOWLEDGMENTS

We thank Núria Bonet and Raquel Rasal (Servei de Genòmica UPF) for their help during the library preparation and sequencing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01315/full#supplementary-material

REFERENCES

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., et al. (2019). Characterizing the major structural variant alleles of the human genome. *Cell* 176, 663–675.e19. doi: 10.1016/j.cell.2018.12.019

Collins, R. L., Brand, H., Redin, C. E., Hanscom, C., Antolik, C., Stone, M. R., et al. (2017). Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* 18, 36. doi: 10.1186/s13059-017-1158-6

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* 464, 704–712. doi: 10.1038/nature08516

Gabrieli, T., Sharim, H., Fridman, D., Arbib, N., Michaeli, Y., and Ebenstein, Y. (2018). Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH). *Nucleic Acids Res.* 46, e87. doi: 10.1093/nar/gky411

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., et al. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci. Rep.* 7, 1–10. doi: 10.1038/s41598-017-03996-z

Gribble, S. M., Ng, B. L., Prigmore, E., Fitzgerald, T., and Carter, N. P. (2009). Array painting: a protocol for the rapid analysis of aberrant chromosomes using DNA microarrays. *Nat. Protoc.* 4, 1722–1736. doi: 10.1038/nprot.2009.183

Huddleston, J., and Eichler, E. E. (2016). An incomplete understanding of human genetic variation. *Genetics* 202, 1251–1254. doi: 10.1534/genetics.115.180539

Jain, M., Olsen, H. E., Turner, D. J., Stoddart, D., Bulazel, K. V., Paten, B., et al. (2018). Linear assembly of a human centromere on the y chromosome. *Nat. Biotechnol.* 36, 321–323. doi: 10.1038/nbt.4109

Jiang, W., Zhao, X., Gabrieli, T., Lou, C., Ebenstein, Y., and Zhu, T. F. (2015). Cas9-assisted targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nat. Commun.* 6, 8101. doi: 10.1038/ncomms9101

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly *via* adaptive κ-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Kozarewa, I., Armisen, J., Gardner, A. F., Slatko, B. E., and Hendrickson, C. L. (2015). Overview of target enrichment strategies. *Curr. Protoc. Mol. Biol.* 112, 7.21.1–7.2123. doi: 10.1002/0471142727.mb0721s112

Kuderna, L. F. K., Lizano, E., Julià, E., Gomez-Garrido, J., Serres-Armero, A., Kuhlwilm, M., et al. (2019). Selective single molecule sequencing and assembly of a human Y chromosome of African origin. *Nat. Commun.* 10, 4. doi: 10.1038/s41467-018-07885-5.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12. doi: 10.1186/gb-2004-5-2-r12

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature* 470, 59–65. doi: 10.1038/nature09708

Nattestad, M., and Schatz, M. C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32, 3021–3023. doi: 10.1093/bioinformatics/btw369

Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2019). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* 35, 2193–2198. doi: 10.1093/bioinformatics/bty841

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., et al. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* 15, 461–468. doi: 10.1038/s41592-018-0001-7

Seo, J. S., Rhie, A., Kim, J., Lee, S., Sohn, M. H., Kim, C. U., et al. (2016). De novo assembly and phasing of a Korean human genome. *Nature* 538, 243–247. doi: 10.1038/nature20098

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., et al. (2016). Long-read sequencing and *de novo* assembly of a Chinese genome. *Nat. Commun.* 7, 12065. doi: 10.1038/ncomms12065

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* 14, 407–410. doi: 10.1038/nmeth.4184

Smit, A., Hubley, R., and Green, P. (2013–2015). RepeatMasker Open-4.0. http://www.repeatmasker.org.

Stephens, Z., Wang, C., Iyer, R. K., and Kocher, J.-P. (2018). Detection and visualization of complex structural variants from long reads. *BMC Bioinf.* 19, 508. doi: 10.1186/s12859-018-2539-x

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81. doi: 10.1038/nature15394

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* 14, 125–138. doi: 10.1038/nrg3373

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter 4

Structural variant at chromosome 1q44 involving the *NLRP3* gene as a novel mechanism causing cryopyrin-associated periodic syndromes

Sukru Cekic[1,*], Manuel Solís-Moruno[2,*], Alejandro Peñín-Franch[3,*], Anna Mensa-Vilaró[4], Núria Bonet[2], María del Mar Rodríguez-Rivera[5,6], Carme Melero[5,6], Marta Bódalo[7], Roger Anglada[2], María José Maleno[8], Laura Hurtado-Navarro[3], Cristina Molina-López[3], Susana Plaza[4], Virginia Fabregat[4], Jordi Yagüe[4,8,9], Diego Angosto-Bazarra[3], Blanca Espinet[5,6], Lara Nonell[7], Montse Plana[8], Oscar Fornas[10], Ferran Casals[2,#], Pablo Pelegrin[3,11,#], Sara Sebnem Kilic[1,#], Juan I. Arostegui[4,8,9,#]

[*] These authors contributed equally to this work.

[#] These authors jointly directed this work.

Work in progress. Manuscript in preparation.

Abstract

Introduction

Cryopyrin-associated periodic syndromes (CAPS) are a consequence of monoallelic *gain-of-function NLRP3* mutations that lead to a hyperactive NLRP3 inflammasome, and are extremely sensitive to anti-interleukin 1 (IL-1) drugs.

Objective

To elucidate the molecular basis of a dominantly-inherited disease observed in a family with clinical features and outcome to anti-IL-1 drugs similar to CAPS.

Material and methods

Genetic studies were performed using Sanger sequencing, whole exome sequencing (WES), SNPs genotyping, fluorescence *in situ* hybridization and genomic sequencing. Patients' chromosome 1 was physically isolated using flow cytometry from EBV-immortalized B cells. *In vitro* analyses will be performed to assess the functional consequences of the detected variant.

Results

Sanger sequencing of *NLRP3* gene and WES did not yield positive results. Subsequent analyses using high-density SNPs genotyping detected a ~1 Mb genomic duplication on chromosome 1q44 including the *NLRP3* gene as perfectly segregating with the disease in this family. The sequencing of patients' chromosome 1 elucidated the organization of this novel structural variant, revealing the presence of an extra copy of *NLRP3* (exons 1-8). *In vitro* analyses will be performed to show if this structural variant provokes a NLRP3 inflammasome hyperactivation similar to the missense *NLRP3* mutations previously described as causing CAPS.

Conclusions

We describe for the first time a genomic structural variant involving the *NLRP3* gene as a novel mechanism causing CAPS. Consequently, this kind of gene variants should be considered in future strategies of disease diagnosis. As they are often lost by either Sanger and next generation sequencing, alternative methods of genetic analyses should be employed in selected candidate patients.

Introduction

Cryopyrin associated periodic syndromes (CAPS) are a set of diseases caused by gain-of-function mutations in the *NLRP3* gene (Hoffman et al. 2001), located at chromosome 1q44, all of them with autosomal dominant inheritance. This gene encodes the cryopyrin protein, which is the main component of the NLRP3 inflammasome. CAPS are diseases caused by an hyperactivation of the activity of the NLRP3 inflammasome, which generates an increased activity in the proinflammatory cytokines IL-1β and IL-18. This way, the treatment of choice for CAPS patients are usually blockers of the IL-1 activity, such as anakinra, canakinumab or rilonacept (Lachmann et al. 2009). A prompt clinical response is accomplished when the treatment is given, and the systemic inflammatory symptoms observed in these patients –fever, urticaria– are fast cleared.

In this work, we present the case of a family of Turkish ancestry with at least five generations presenting cases of an autoinflammatory disease. Symptoms were compatible with the Muckle-Wells syndrome (Muckle and Wells 1962), an intermediate form of CAPS. We performed several different genetic studies in 13 members of the family across the last three generations, 11 patients and two healthy individuals. We performed whole exome sequencing in samples coming from five individuals, four patients and one healthy. Besides, we isolated by flow cytometry the chromosome 1 of two patients. We discovered and characterized a ~1 Mb duplication encompassing several genes, with *NLRP3* included –except from the last exon–. This region also showed two inversion events. We propose that this structural variant (SV) is related to the phenotype described in this family, since it is present in all cases and in none of the healthy individuals, segregating perfectly with the disease.

Material and methods

Family

We explored the case of a family from Turkey with an autoinflammatory disease compatible with CAPS. Five different generations presented the disease, with no discrimination between males and females (Figure 1), pointing to an autosomal dominant inheritance. The Ethical Review Board of the Hospital Clínic, Barcelona, approved the study. All investigations were performed in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments.

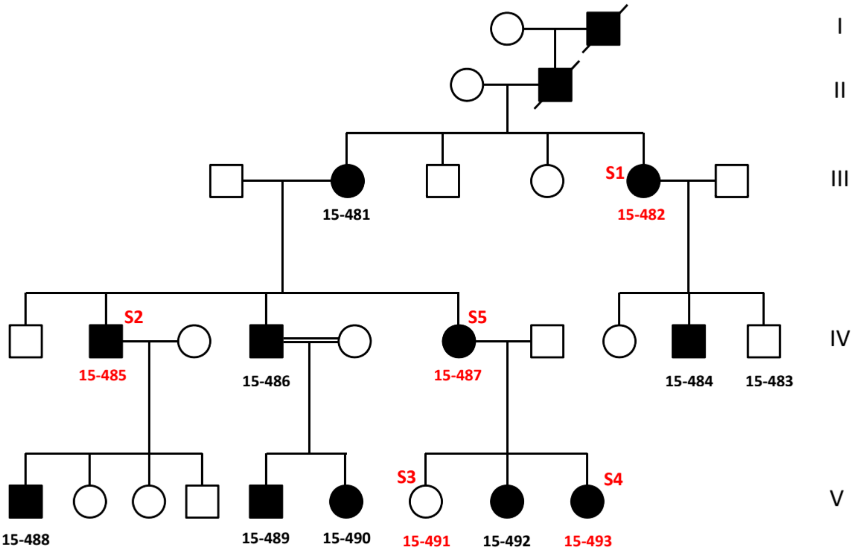


Figure 1. Pedigree of the analysed family. Five different generations have shown autoinflammatory symptoms.

Samples and sequencing

We extracted DNA from peripheral blood for 13 individuals of the family, marked 15-481 to 15-493 in Figure 1, in order to perform different genetic studies. For individuals 15-482, 15-485, 15-487, 15-491 and 15-493 (also marked S1 to S5) whole exome sequencing was performed. The libraries were sequenced in a NextSeq Illumina platform in High Output 2 × 150 paired-end cycles runs to a mean coverage of ~60X.

Mapping and variant calling

We mapped our samples to the hg19 human reference genome (UCSC) using BWA-MEM (Li 2013) algorithm (v. 0.7.16a-r1181). We marked duplicated reads with picard tools (v. 1.93), realigned indels and performed base quality score recalibration with the correspondent tools from GATK (McKenna et al. 2010) (v. 3.4-46-gbc02625). We also used GATK's HaplotypeCaller to call germline SNVs and indels.

To annotate the genetic variants, we used SnpEff (Cingolani, Platts, et al. 2012) (v. 4.3i) and, to incorporate SIFT (Vaser et al. 2016) and PolyPhen-2 (Adzhubei et al. 2010) predictions, as well as gnomAD (Karczewski et al. 2020) population frequencies, we used SnpSift (Cingolani, Patel, et al. 2012) (v. 4.3i).

For the structural variant calling, we used XHMM (Fromer et al. 2012) (v1.0) and CoNIFER (Krumm et al. 2012) (v0.2.2). For a better performance of both software, we also added another 27 samples generated following the same protocol that were used internally by the programs to correct errors. We used XHMM with default parameters and CoNIFER with the following ones:

```
python $conifer rpkm --probes $probes --input
$input/BQSR_S${i}.bam --output $input/BQSR_all.rpkm.txt
```

```
python $conifer analyze --probes $probes --rpkm_dir $RPKMs
--output $results/analysis_all.hdf5 --svd 4 --write_svals
$results/singular_values_all.txt --plot_scree
$results/screenplot_all.png --write_sd $results/sd_values_all.txt
```

```
python $conifer call --input $results/analysis_call.hdf5 --output
$results/call_BQSR.txt
```

CytoScan HD array

We used the CytoScan HD array (Affymetrix, Santa Clara, California, USA) to explore the copy number variant landscape and the possible runs of homozygosity of the 13 individuals marked 15-481 to 15-493. These analyses were performed in the Microarray Service of the Instituto Hospital del Mar de Investigaciones Médicas (IMIM).

TaqMan Copy Number Assay

To explore the copy number state of the *NLRP3* gene, we used the TaqMan Copy Number Assay (Thermo Fisher Scientific, Waltham, Massachusetts, USA). We ordered probes for exons 3, 5, 7, 8 and 9 (NM_001243133.1 transcript).

FISH

We derived lymphoblastoid cell lines from individuals 15-490 and 15-481. We used them to perform fluorescent *in situ* hybridization (FISH) in order to test where the duplicated region was located. We designed three probes that hybridized to the region encompassing genes *AHCTF1*, *ZNF695* and *ZNF124* respectively. These analyses were performed in the Laboratory of Molecular Cytogenetic of the Hospital del Mar.

Nanopore sequencing, assembly and structural variant calling

From the two derived lymphoblastoid cell lines, we isolated and enriched by flow cytometry, as described in Chapter 3 of this thesis, the chromosome 1 of the affected individuals 15-490 and 15-481. We sequenced each sample in two different runs of a MinION device (Oxford Nanopore Technologies). For the basecalling, we used guppy (v. 3.2.2) with the following parameters:

```
guppy_basecaller --compress_fastq -i $input -s $output
--cpu_threads_per_caller 4 --num_callers 1 --flowcell FLO-MIN106
--kit SQK-LSK109
```

We merged the fastq files obtained for each sample and mapped the raw reads with Minimap2 (Li 2018) (v. 2.1) and NGMLR (Sedlazeck et al. 2018) (v. 0.2.6), using default parameters, to the human reference genome hg38 (UCSC). In Figure 2 we show the number of reads mapping to each chromosome for each individual and run after using Minimap2. We used mosdepth (Pedersen and Quinlan 2018) (v. 0.2.3) to calculate the resulting mean coverage in chromosome 1, that reached ~20X for both individuals.

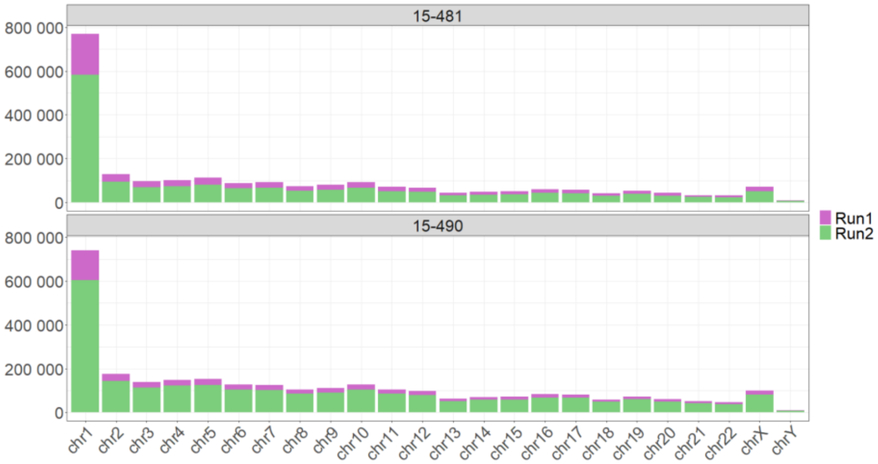


Figure 2. Number of reads mapping to each chromosome per individual and run. We observe the enrichment in reads in chromosome 1 in both cases.

We performed structural variant calling using Sniffles (Sedlazeck et al. 2018) (v. 1.0.11) and SVIM (Heller and Vingron 2019) with default parameters.

We also aimed to compare our data to the chromosome 1 of the human reference genome hg38 (UCSC). For that, we extracted all reads mapping to chromosome 1 in our data and we assembled them using Canu (Koren et al. 2017) (v. 1.8). For the comparison purpose, we aligned the produced assemblies to the chromosome 1 of hg38 by using the nucmer tool in MUMmer (Eisen et al. 2000) (v 3.9.4). We filtered our data with delta-filter, also included in the MUMmer package. We used the following code:

```
canu -p ${i}_${a}-chr1 -d $output genomeSize=250m -nanopore-raw
$input/${i}_${a}_reads_chr1.fastq.gz
```

```
nucmer --maxmatch -l 100 -c 500 $hg38_chr1 $input/${i}_${a}-
chr1.contigs.fasta --prefix $output/${i}_${a}_vs.hg38_chr1
```

```
delta-filter -i 80 -l 1000 $input/${i}_${a}_vs.hg38_chr1.delta
```

We polished our data using Nanopolish (Loman, Quick, and Simpson 2015) (v. 0.11.1) with the following code:

```
python $makerange --segment-length 500000 --overlap-length 1000
$input/${i}_${a}_Minimap2/${i}_${a}_Minimap2-chr1.contigs.fasta |
$parallel $srun nanopolish variants --consensus
-o $output/polished_${i}_${a}.{1}.vcf -w {1}
-r $fastq/${i}_${a}_reads_chr1.fastq.gz
-b $input/${i}_${a}_mapped2assembly_sorted.bam
-g $input/${i}_${a}/${i}_${a}_chr1.contigs.fasta --min-candidate-
frequency 0.1 -t 8
```

We introduced the corrections in the assemblies with vcf2fasta tool of Nanopolish:

```
nanopolish vcf2fasta -g $assembly/${i}_${a}_chr1.contigs.fasta
$calls/polished_${i}_${a}.vcf >
$assembly/${i}_${a}_chr1.contigs_polished.fasta
```

We finally plotted our results using mummerplot tool from MUMmer (Figure 3):

```
mummerplot -R $hg38_chr1 -Q $assembly/${i}_${a}-
chr1.contigs_polished.fasta --coverage --filter --layout -t png –p
$output/${i}_${a}_plot_polished
${i}_${a}_vs.hg38_chr1_polished.delta
```

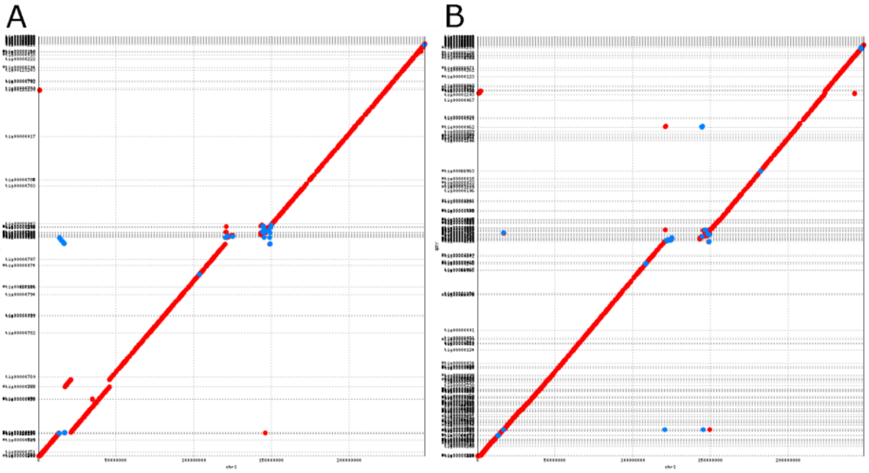


Figure 3. Dot plot with the results of our polished assemblies and chromosome 1 of hg38 reference genome. **A.** 15-490 individual. **B.** 15-481 individual.

Using the polished assembly, we run Assemblytics (Nattestad and Schatz 2016) to call structural variants on our assembly in comparison to the chromosome 1 of the human reference genome. We looked for large variants of 10,000-1,500,000 bp long. We also used SVIM-asm (Heller and Vingron 2020), an extension of the SVIM software for assembled data. This way, we mapped our assemblies to the human reference genome, as described in the documentation, and then we run the software with default parameters.

Results

CytoScan HD array

We discovered a ~1 Mb duplication in the final part of chromosome 1, 1q44, in all individuals with autoinflammatory symptoms. This duplication was not present in healthy individuals 15-483 and 15-491. The copy number state of all individuals suffering from the disease was three, except for 15-490, who presented four copies of the duplicated region. The ~1 Mb region harboured the following genes: *SMYD3*, *TFB2M*, *CNST*, *AHCTF1*, *ZNF695*, *ZNF124*, *ZNF496* and *NLRP3*.

Variant calling from the whole exome sequencing data

We found no significant SNVs or indels that could be linked to the autoinflammatory disease. When exploring the structural variant landscape of individuals S1-S5, no significant results were found when using XHMM. With CoNIFER, we found signals of duplication around the final part of the ~1 Mb region detected by CytoScan HD, where the *NLRP3* gene is located (Figure 4).

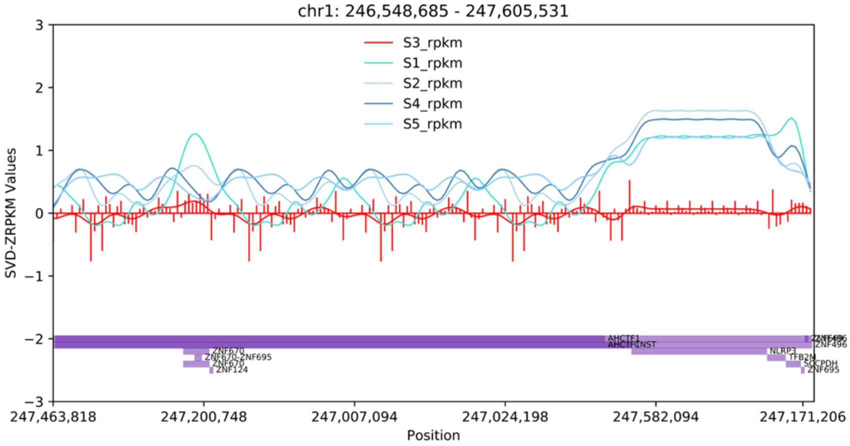


Figure 4. CoNIFER plot. We show the ~1 Mb area that appeared as duplicated according to CythoScan HD array. The only signal called as duplicated in the whole exome sequencing data was located in the final part, encompassing the *NLRP3* gene. S3 individual (15-491), which was healthy, did not present this signal.

TaqMan Copy Number Assay

We wanted to further explore the copy number state of the *NLRP3* gene and, for that purpose, we used the TaqMan Copy Number Assay with probes for exons 3, 5, 7, 8 and 9 (the last one). We found that all individuals suffering from the disease presented three copies of all exons and two copies of the final one (Figure 5). Individual 15-490 presented four copies of all exons and two copies of exon 9. The healthy individual 15-491 was confirmed to have two copies of the whole gene.

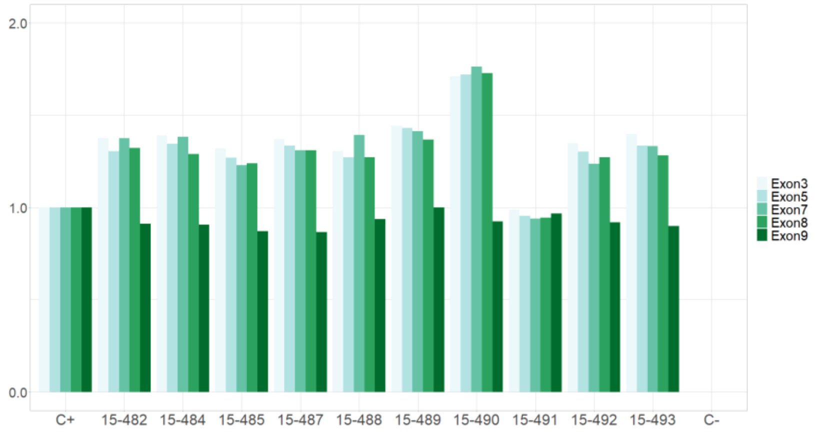


Figure 5. TaqMan Copy Number Assay results. A value of 1 in the y axis indicates two copies of the explored exon, while a value of 2 indicates four copies.

FISH

We performed FISH using cell lines generated from individuals 15-490 and 15-481, the first one having four copies of the duplicated region and the second one having three. With this experiment, we confirmed that the duplication event was located within the same chromosome 1 in both cases (Figure 6).

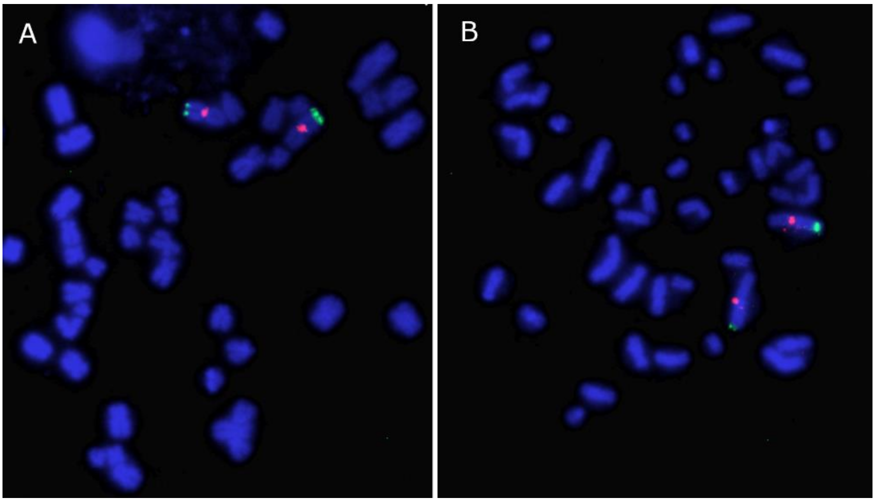


Figure 6. FISH results. The centromere of chromosome 1 is marked in pink and, in green, we observe the designed probes hybridizing to our region of interest. We see that all signals come from chromosome 1. **A.** 15-490 individual. **B.** 15-481 individual.

Nanopore structural variant calling

We used NanoPlot (v. 1.14.1) to obtain several stats of the output data:

First run of individual 15-490: Total number of reads: 579,087, mean read length: 5,218.9, median read length: 2,662.0, read length N50: 15,250. Second run of individual 15-490: Total number of reads: 1,758,376, mean read length: 4,901.5, median read length: 1,745.0, read length N50: 16,960. First run of individual 15-481: Total number of reads: 490,217, mean read length: 5,055.4, median read length: 888.0, read length N50: 27,338. Second run of individual 15-481: Total number of reads: 2,354,777, mean read length: 4,323.0, median read length: 1,926.0, read length N50: 11,520.

We visually inspected the region around the *NLRP3* gene in both individuals using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011), plus the Chinese individual of Chapter 3 for comparison purposes (Figure 7).

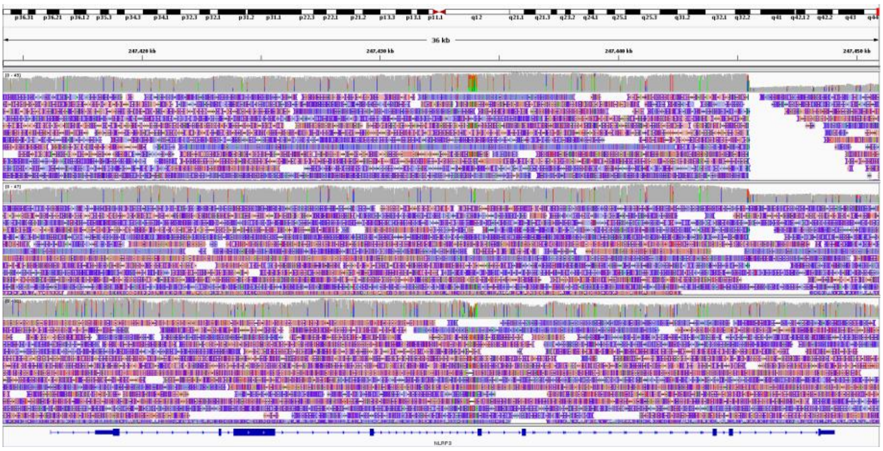


Figure 7. IGV screenshot of the *NLRP3* locus. Top individual is 15-490, middle individual is 15-481 and bottom individual is the one from Chapter 3. A sudden drop of coverage in both Turkish individuals can be observed before the final exon of *NLRP3*, while the coverage is homogenous in the Chinese individual.

We were not able to call any duplication in our region of interest when using Sniffles, SVIM, Assemblytics and SVIM-asm, although it was previously confirmed with CythoScan HD, TaqMan Copy Number Assay and the whole exome sequencing data. However, as mentioned before, the mean coverage of our samples is ~20X while, in this 1 Mb region, it reaches 38.5X in 15-490 and 33.3X in 15-481, what is in concordance with the four and three copies respectively found with CythoScan HD and TaqMan Copy Number Assay.

Interestingly, two different inversion events were called in this ~1 Mb region after mapping with NGMLR and calling structural variants with SVIM. In individual 15-490 it was called an inversion between positions chr1:246,384,518-246,940,676 (*SMYD3 – AHCTF1*) and another one in chr1:246,919,601-247,445,324 (*AHCTF1 – NLRP3*). This shows an overlap between positions chr1:246,919,601-246,940,676, falling between the end of *AHCTF1* and the intergenic region between this gene and *ZNF695*. With the same strategy, in individual 15-481, we found one inversion in chr1:246,384,584-246,940,610 (*SMYD3 – AHCTF1*) and another one in chr1:246,777,162-247,637,264 (*SCCPDH – OR2G3*). The overlap is between chr1:246,777,162-246,940,610, encompassing, once again, the *AHCTF1* gene.

After mapping with Minimap2 these inversions are not called in 15-481 but they are in 15-490. The genomic coordinates are now chr1:246,384,402-246,940,794 and chr1:246,777,162-247,637,264. The second one is called exactly in the same positions. The region between both events falls between positions chr1:246,919,724-246,940,509 (Figure 8).

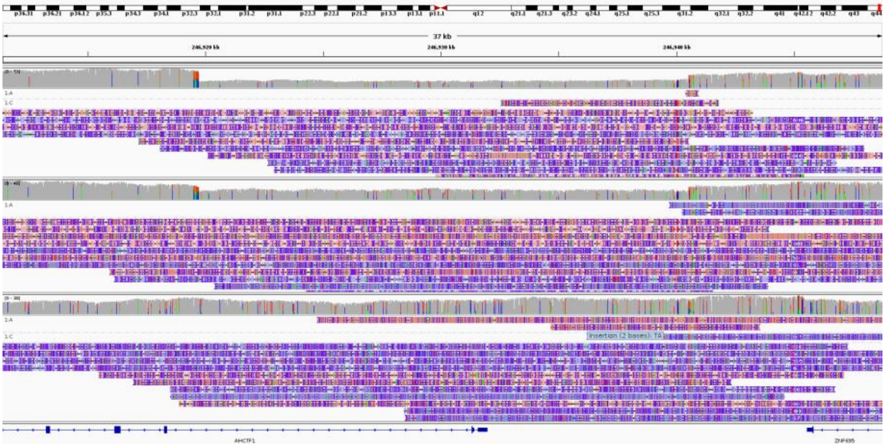


Figure 8. IGV screenshot of the region between the inversion events. Top individual is 15-490, middle individual is 15-481 and bottom individual is the one from Chapter 3. The valley cannot be observed in the Chinese individual.

When using SVIM-asm, the inversion between positions chr1:246,919,496-247,445,435 (from *AHCTF1* to *NLRP3*) in individual 15-490 was again called.

Analysis of the reads with secondary alignment

Secondary alignment (SA) is the term used to define those reads, in a given sequencing experiment, that also map to a different region than the primary position given by the mapping software.

We found that both individuals, 15-490 (Supplementary Table 1) and 15-481 (Supplementary Table 2), presented reads with SA at the beginning and at the end of both duplication events. Besides, all reads with SA map to the opposite strand in comparison to their primary alignment, i. e., if the primary read mapped in the positive strand, its SA mapped in the negative strand. Interestingly, the mapping quality of all SA is MQ=60, which means it is perfect. All reads located at the breakpoints of the two inversion events mapped to the valley area between them.

Discussion

Taken together, our results suggest that, in the ~1 Mb area reported by CytoScan HD array results, there are two different inversion events that are also duplicated. These events would exclude the final exon of the *NLRP3* gene. Cryopyrin protein has been reported to be functional without the final leucine-rich repeat (LRR) domain (Hafner-Bratkovič et al. 2018). We suggest a novel genetic model underlying CAPS in this Turkish family, in which there are no missense mutations causing a gain-of-function effect. In this case, an extra copy of the *NLRP3* gene would be sufficient to cause disease in a genetic dose-dependent manner. We have observed an extra copy of *NLRP3* in all individuals presenting the disease (two extra copies in the case of 15-490), with healthy individuals showing no gain of genetic material.

NLRP3 somatic variants with moderate variant allele frequencies in whole blood (4.90%, 5.08%) have been reported to cause CAPS (Mensa-Vilaró et al. 2018). This way, if there are no regulatory mechanisms controlling the production of cryopyrin, an extra copy of *NLRP3* seems to be sufficient to cause disease. To assure that this scenario is actually occurring, further functional studies will be performed. This way, we would be able to characterize the activity of an extra copy of the *NLRP3* gene *in vitro*.

The affected individuals of this family were treated with anakinra, which controlled their symptomatology.

Bibliography

Adzhubei, Ivan A., Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. "A Method and Server for Predicting Damaging Missense Mutations." *Nature Methods* 7 (4): 248–49. https://doi.org/10.1038/nmeth0410-248.

Cingolani, Pablo, Viral M. Patel, Melissa Coon, Tung Nguyen, Susan J. Land, Douglas M. Ruden, and Xiangyi Lu. 2012. "Using Drosophila Melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift."

Frontiers in Genetics 3 (MAR): 1–9. https://doi.org/10.3389/fgene.2012.00035.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff." *Fly* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

Eisen, J. A., J. F. Heidelberg, O. White, and S. L. Salzberg. 2000. "Evidence for Symmetric Chromosomal Inversions around the Replication Origin in Bacteria." *Genome Biology* 1 (6): 1–9. https://doi.org/10.1186/gb-2000-1-6-research0011.

Fromer, Menachem, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker, et al. 2012. "Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth." *American Journal of Human Genetics* 91 (4): 597–607. https://doi.org/10.1016/j.ajhg.2012.08.005.

Hafner-Bratkovič, Iva, Petra Sušjan, Duško Lainšček, Ana Tapia-Abellán, Kosta Cerović, Lucija Kadunc, Diego Angosto-Bazarra, Pablo Pelegrin, and Roman Jerala. 2018. "NLRP3 Lacking the Leucine-Rich Repeat Domain Can Be Fully Activated via the Canonical Inflammasome Pathway." *Nature Communications* 9 (1): 5182. https://doi.org/10.1038/s41467-018-07573-4.

Heller, David, and Martin Vingron. 2019. "SVIM: Structural Variant Identification Using Mapped Long Reads." *Bioinformatics* 35 (17): 2907–15. https://doi.org/10.1093/bioinformatics/btz041.

———. 2020. "SVIM-Asm: Structural Variant Detection from Haploid and Diploid Genome Assemblies." *Bioinformatics*, no. December: 1–3. https://doi.org/10.1093/bioinformatics/btaa1034.

Hoffman, Hal M., James L. Mueller, David H. Broide, Alan A. Wanderer, and Richard D. Kolodner. 2001. "Mutation of a New Gene Encoding a Putative Pyrin-like Protein Causes Familial Cold Autoinflammatory Syndrome and Muckle-Wells Syndrome." *Nature Genetics* 29 (3): 301–5. https://doi.org/10.1038/ng756.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581 (7809): 434–43. https://doi.org/10.1038/s41586-020-2308-7.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive *κ*-Mer Weighting and Repeat Separation." *Genome Research* 27 (5): 722–36. https://doi.org/10.1101/gr.215087.116.

Krumm, Niklas, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, Aaron R. Quinlan, Deborah A. Nickerson, and Evan E. Eichler. 2012. "Copy Number Variation Detection and Genotyping from Exome Sequence Data." *Genome Research* 22 (8): 1525–32. https://doi.org/10.1101/gr.138115.112.

Lachmann, Helen J., Isabelle Kone-Paut, Jasmin B. Kuemmerle-Deschner, Kieron S. Leslie, Eric Hachulla, Pierre Quartier, Xavier Gitton, Albert Widmer, Neha Patel, and Philip N. Hawkins. 2009. "Use of Canakinumab in the Cryopyrin-Associated Periodic Syndrome." *New England Journal of Medicine* 360 (23): 2416–25. https://doi.org/10.1056/nejmoa0810787.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *ArXiv:1303.3997 [q-Bio.GN]*, March. http://arxiv.org/abs/1303.3997.

———. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Loman, Nicholas J., Joshua Quick, and Jared T. Simpson. 2015. "A Complete Bacterial Genome Assembled de Novo Using Only Nanopore Sequencing Data." *Nature Methods* 12 (8): 733–35. https://doi.org/10.1038/nmeth.3444.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Mensa-Vilaró, Anna, María Bravo García-Morato, Oscar de la Calle-Martin, Clara Franco-Jarava, María Teresa Martínez-Saavedra, Luis I. González-Granado, Eva González-Roca, et al. 2018. "Unexpected Relevant Role of Gene Mosaicism in Primary Immunodeficiency Diseases." *Journal of Allergy and Clinical Immunology*, 1–10. https://doi.org/10.1016/j.jaci.2018.09.009.

Muckle, Thomas J., and Michael Wells. 1962. "URTICARIA, DEAFNESS, AND AMYLOIDOSIS: A NEW HEREDO-FAMILIAL SYNDROME." *QJM: An International Journal of Medicine* 31 (2): 235–48. https://doi.org/10.1093/oxfordjournals.qjmed.a066967.

Nattestad, Maria, and Michael C. Schatz. 2016. "Assemblytics: A Web Analytics Tool for the Detection of Variants from an Assembly." *Bioinformatics* 32 (19): 3021–23. https://doi.org/10.1093/bioinformatics/btw369.

Pedersen, Brent S., and Aaron R. Quinlan. 2018. "Mosdepth: Quick Coverage Calculation for Genomes and Exomes." Edited by John Hancock. *Bioinformatics* 34 (5): 867–68. https://doi.org/10.1093/bioinformatics/btx699.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. "Integrative Genomics Viewer." *Nature Biotechnology* 29 (1): 24–26. https://doi.org/10.1038/nbt.1754.

Sedlazeck, Fritz J., Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. 2018. "Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing." *Nature Methods* 15 (6): 461–68. https://doi.org/10.1038/s41592-018-0001-7.

Vaser, Robert, Swarnaseetha Adusumalli, Sim Ngak Leng, Mile Sikic, and Pauline C. Ng. 2016. "SIFT Missense Predictions for Genomes." *Nature Protocols* 11 (1): 1–9. https://doi.org/10.1038/nprot.2015.123.

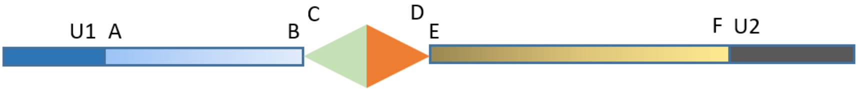
Supplementary Table 1. Breakpoints of the two inversion events in the ~1 Mb area in individual 15-490.

Position	Name	Gene	#Total	#No SA	#SA	%SA	Position SA	Name SA
246384401	U1	*SMYD3*	21	21	0	0%	—	—
246384402	A	*SMYD3*	33	22	11	33.33%	246940510	E
246919723	B	*AHCTF1*	40	22	18	45%	~247441254	F
246919724	C	*AHCTF1*	21	21	0	0%	—	—
246940509	D	Intergenic	27	26	1*	0%	(*149591775)	(*—)
246940510	E	Intergenic	39	27	12	30.77%	246384402	A
247445514	F	*NLRP3*	26	12	14	53.85%	~246916166	B
247445515	U2	*NLRP3*	11	11	0	0%	—	—

Position refers to the position of chr1 in hg38 coordinates. *One read maps to another far position of chr1, which does not seem to be related to these events. # refers to number of reads. For the column Name, please see Supplementary Figure 1.

Supplementary Table 2. Breakpoints of the two inversion events in the ~1 Mb area in individual 15-481.

Position	Name	Gene	#Total	#No SA	#SA	%SA	Position SA	Name SA
246384401	U1	*SMYD3*	17	17	0	0%	—	—
246384402	A	*SMYD3*	27	17	10	37.04%	246940510	E
246919723	B	*AHCTF1*	31	19	12	38.71%	~247441254	F
246919724	C	*AHCTF1*	22	19	3	13.67%	2: ~246885970; 1: 247417837	C;F
246940509	D	Intergenic	22	21	1*	0%	(*21317611)	(*—)
246940510	E	Intergenic	30	22	9*	26.67%	246384402 (*21317611)	A (*—)
247445514	F	*NLRP3*	27	21	6	22.22%	~246916166	B
247445515	U2	*NLRP3*	22	21	1	4.55%	247458889	U2



Supplementary Figure 1. Graphical representation of the ~1 Mb duplicated according to CytoScan HD results. **U1** is the first position outside of the first inversion event, **A** is the first position within the first inversion event (*SMYD3*), **B** is the last position within the inversion event (*AHCTF1*), **C** is the first position of the region between the two inversion events (*AHCTF1*), **D** is the last position of the region between the two inversion events, **E** is the first position within the second inversion event, **F** is the last position within the second inversion event (*NLRP3*) and **U2** is the first position outside of the second inversion event.

Discussion

Summary of the results

In this thesis, I have explored the ability of MPS data to detect somatic and germline SNVs and SVs in the context of autoinflammatory diseases, and also somatic SNVs in other PIDs.

I have demonstrated the power of high coverage WES to call somatic SNVs from DNA extracted from whole blood, oral mucosa and urine. Besides, I have taken advantage of the generated dataset to further explore the load of somatic coding variants in whole blood, which resulted to be low in most cases.

I have also proven the capacity of MPS, both from short Illumina reads and long Nanopore reads, to uncover pathogenic SVs in two genes, with different inheritance models, in two distinct families; all in combination with different laboratory techniques. I have identified a novel deletion in *IL1RN*, transmitted from the mother to two siblings affected with DIRA. The father also carried a novel variant discovered by our collaborators. In his case, he presented a SNV affecting a splice site of the same gene. I have also first identified and characterized a novel duplication spanning approximately 1 Mb, and including the *NLRP3* gene –except from the last exon– in a family with several cases of CAPS. Additionally, I have participated in the isolation with flow cytometry of the chromosome 1 of a Chinese individual from the 1000 Genomes Project, which was sequenced using an ONT MinION device. This was a proof of concept that enabled the study of SVs focusing on this chromosome, in which *NLRP3* is located, with no need to sequence the whole genome.

Somatic variation in autoinflammatory and other diseases

Up to six different genes have been reported to harbour causal somatic variants of autoinflammatory diseases: *MEFV*, *NLRC4*, *NLRP3*, *NOD2*, *TMEM173* and *TNFRSF1A*. Since somatic variants display lower VAF than germline variants in a given tissue, it is not surprising that all diseases linked to mutations in those genes present an autosomal dominant inheritance. As mentioned in the Introduction of this thesis, autosomal dominant diseases caused by gain of function mutations are not rare in autoinflammatory disorders (Manthiram et al. 2017), and it is likely that new somatic causing variants will be discovered in coming years. All Mendelian diseases with a dominant inheritance can potentially be caused by somatic variants. Causing a recessive disease with only somatic variants would be more difficult, since two different early co-occurring mutational events would be needed. Interestingly, a case of one heterozygous variant combined with a somatic one in *TNFRSF6* has been reported to be causing the recessive autoimmune lymphoproliferative syndrome (Magerus-Chatinet et al. 2011).

In our first work we sequenced, to a high coverage of ~245X, the exome of 16 samples from 11 individuals with already known pathogenic somatic variants related not only to autoinflammatory diseases but also to other PIDs. Our aim was (1) to find those variants in our sequencing data, as well as (2) to take advantage of the generated dataset to explore the somatic variant landscape of these individuals in whole blood.

For the first purpose, all somatic causing variants but one were called in our experiment. The variant that was not called was not present in any of the produced reads, so it was technically impossible to retrieve it. Probably, increasing the sequencing coverage would have solved this issue. Coverage is a common concern in sequencing experiments, especially when trying to discover somatic variants in bulk tissue, as they can be diluted among sequencing and other errors. We implemented a set of stringent filters to tackle this issue but, in the end, the use of a set of candidate genes (the IUIS list of genes related to PIDs) was the step that reduced the most the number of candidate variants. Without a panel of candidate genes, more stringent filters can eliminate true somatic variants. However, this also comes with a drawback which is that, if such panels are always used, no new genes related to disease can be discovered.

With these results, we show the importance of taking into account somatic variants in future genetic studies. This is especially relevant when working with autoinflammatory and other haematological diseases, in which the affected tissue –blood– is easily sampled. Besides, it has been proven that the reanalysis of MPS data enhances diagnostic rates (Costain et al. 2018; Sun et al. 2019; Won et al. 2020; Wright et al. 2018) since novel knowledge and tools are constantly generated. This way, we also suggest reanalysing previously generated data considering the somatic model.

When exploring the somatic variant landscape of these individuals, only the previously known pathogenic somatic variant, plus one in some cases, was found in their whole exomes. Our validation rate was modest, since we generated, from the WES blood samples, a list of 461 somatic candidate variants across the 11 individuals. Only 34/461 variants were validated by amplicon-based deep sequencing. One individual presented a total of 17 validated somatic variants, all of which clustered in three different groups of VAFs, suggesting three different clonal events in the context of clonal haematopoiesis. Among the 461 candidate variants, 30 were found in zinc finger proteins, 20 of them in chromosome 19, and none was validated. These genes (*ZNF136*, *ZNF264*, *ZNF304*, *ZNF320*…) are an example of how regions with high homology can confound the somatic variant discovery process, even with a set of highly stringent filters.

The role of somatic variants in healthy tissue and diseases other than cancer is being recently explored. Rather than using bulk tissue, the field is shifting towards the use of microdissections (Abascal et al. 2021) and single cell sequencing (Xing et al. 2021). These strategies allow study from one cell to few hundreds, which helps discriminating somatic from germline variants more easily.

Recent works have shown different numbers of somatic SNVs in a whole genome: 200-2,000 depending on age in oesophageal epithelium (Martincorena et al. 2018), around 40 in exomes and 1,880 in genomes in urothelium (Lawson et al. 2020), 1,000-,1500 in livers of middle aged individuals (Brunner et al. 2019), 1,500-15,000 in colonic crypts (Lee-Six et al. 2019), 210-2,800 in endometrial epithelium (Moore et al. 2020), around 1,500 in neurons (Lodato et al. 2015) and around 790 in exomes of individual melanocytes (Tang et al. 2020). In blood, also correlating with age, 500-1,000 SNVs have been reported in people with ages 30-60 (Osorio et al. 2018). All these numbers reveal one of the main limitations of our study: because of the coverage, we were only able to call SNVs with moderately low VAFs, and thus, all variants with VAF < 1.5% were missed.

Taking our results into account, a mean coverage of around 250X is sufficient to call somatic variants with VAF > 1.5%. This makes this approach suitable when working with medical genomics data, in which variants with moderate frequencies has been observed to cause disease. This happens, for example, in PID (Van Horebeek, Dubois, and Goris 2019) and in other diseases of blood such as acute myeloid leukaemia (Abelson et al. 2018; Desai et al. 2018).

A more detailed explanation addressing the somatic variant landscape in healthy tissue is referred to in the Appendix. We are currently working in a manuscript reviewing this topic.

Detection of newly described pathogenic structural variants

In the second work, we examined the case of a family with four children, two of them with fatal symptoms compatible with DIRA: they exhibited severe, early-onset inflammation affecting mainly skin and bones. We discovered two new variants in the *IL1RN* gene predicted to generate a truncated protein. The father presented the c.318+2T>G variant, while the mother had a 2592 bp deletion extending from the middle of intron one until the middle of exon three of this gene.

A first genetic study with a targeted gene panel for monogenic autoinflammatory diseases revealed in the father and the brother a heterozygous T>G transversion at position +2 of the donor acceptor splicing site of intron 3 of the *IL1RN* gene. This variant was not previously reported in any database and is predicted to impair the normal splicing of this gene. No candidate variants were found in the mother using this strategy.

For the discovery of the aforementioned deletion in the mother, we performed WGS to a mean coverage ~37X. Remarkably, CNVnator did not detect this SV, but we observed a drop of coverage when using IGV to visually examine the *IL1RN* locus. This event, which was not present in two unrelated controls, was compatible with a heterozygous deletion. The potential deletion was then tested and validated by PCR. This case highlights the importance of visual inspection of the mapped reads, since this step can be crucial to identify or discard variants.

The combination of both newly reported variants in two of the children resulted in fatal cases of DIRA. Although the association of the two novel variants and the disease seems clear, we were not able to validate them in the children, since we did not have access to any sample from them.

Deletions generate a loss of genetic material, and they can be related to dominant or recessive inheritance diseases. In the case of dominant inheritance, if one gene is haploinsufficient –a term that refers to those ones that need the two copies to produce the sufficient amount of protein (Fisher and Scambler 1994)– just a heterozygous deletion would be necessary to cause disease. In the case of recessive inheritance, a homozygous deletion or the combination of one deletion plus another variant is needed, as happens in our study with DIRA.

Before the MPS era, there were already 60 diseases linked to deletions in the human genome (Krawczak and Cooper 1991). In a study with 60,000 clinical samples tested for hereditary cancer, only 1.6% (960) of patients presented at least one SV. From the 538 deletions detected, 97.2% (523) turned out to be pathogenic or likely pathogenic, while only 14% (49) of the 350 duplications were classified in these categories (Mu et al. 2019). These numbers remark the pathogenic potential of deletions.

Flow sorting enrichment of individual chromosomes

We discovered an interesting ~1 Mb duplication in chromosome 1 when using CytoScan HD array in a family from Turkey. Several members of this family presented a symptomology and an inheritance pattern compatible with CAPS. This structural event was interesting because the *NLRP3* gene, which is the one mutated in these diseases, was located within the duplicated region. This way, following a previous publication in which chromosome Y was isolated by flow cytometry and sequenced with a ONT MinION device (Kuderna et al. 2019), we aimed to do the same with the biggest human chromosome, chromosome 1. As a proof of concept, we studied SVs in the chromosome 1 of a Chinese individual from the 1000 Genomes Project. We calculated the SVs frequencies in each superpopulation present in the 1000 Genomes Project and, then, we compared them with the two sets of SVs that were generated for our individual.

On the one hand, we assembled the raw reads *de novo* and called SVs in comparison to the human reference genome using Assemblytics (Nattestad and Schatz 2016). On the other hand, we mapped the raw reads to the human reference genome and called SVs with Sniffles (Sedlazeck et al. 2018). We observed a moderate overlap between both call sets, since 61.4% and 56.9% of the SVs were unique to Assemblytics and Sniffles respectively. We found that the SVs of our call set were more frequent in east Asian populations (20.5%), as it was expected from the Chinese ancestry of our individual.

Structural variant in *NLRP3* linked to autoinflammatory disease

Our last work is a collaborative effort from different groups from Turkey and Spain. We examined the case of a family in which, at least, five different generations showed cases of a disease with a symptomology compatible with CAPS. We first examined the WES data of five individuals across three generations, four with the disease and one healthy, for SNVs and indel discovery. No candidate variants were found in this first analysis.

Using the CytoScan HD array, we discovered a duplication in the final region of chromosome 1 (1q44), encompassing around 1 Mb, until the final exon of *NLRP3*. Using the TaqMan Copy Number Assay, we confirmed one extra copy of *NLRP3* in all patients suffering from the disease, two extra copies in one of them and no extra copies in a healthy individual. We also confirmed that the final exon of this gene was not duplicated in any case. We reanalysed the WES data to call SVs and we observed a signal of duplication in the final part of this ~1 Mb region, including precisely the *NLRP3* gene. The duplication was detected with CoNIFER (Krumm et al. 2012), but not with XHMM (Fromer et al. 2012).

After demonstrating that chromosome sorting protocol developed by Kuderna et. al (Kuderna et al. 2019) could be used in chromosome 1 to explore and characterize SVs with long reads, we applied it to two members of this family. Prior to this step, we performed fluorescent *in situ* hybridization to confirm that the duplicated region did not move to another chromosome and that the structural event was, effectively, located in chromosome 1. We then derived lymphoblastoid cell lines from two individuals of this family, one with three copies of *NLRP3* and the other one with four copies. We isolated their chromosome 1 using flow cytometry and sequenced it in an ONT MinION device, achieving a coverage of ~20X in both samples.

Interestingly, the ~1 Mb duplication discovered using CytoScan HD was found to be a complex rearrangement. We observed two different inversion events within this region, both of which contained one or two extra copies depending on the individual. Although the inversion could be called by different software, none of them detected the duplication.

We suggest that this event is the genetic origin of CAPS in this family, considering that it perfectly segregates with the disease. This would be a new genetic model for this disease since, until now, only missense variants causing a gain of function effect were described. In our study, an extra copy of the *NLRP3* gene would be sufficient to generate an aberrant amount of cryopyrin protein that would cause disease. In this sense, cryopyrin has been reported to be functional without the final leucine-rich repeat domain (Hafner-Bratkovič et al. 2018), which is particularly important due to the absence of the final exon of this gene within the duplication.

Overall interpretation

The use and combination of different analytical approaches and techniques allowed us to study and characterize somatic SNVs and germline SVs related to autoinflammatory diseases. From their commercialization, MPS technologies have proven to be a very valuable tool in the discovery of pathogenic variants. However, while very powerful to detect germline SNVs, these methodologies have important limitations to detect other types of variation such as somatic or structural variants. These alternatives and underexplored types of causal genetic variants might explain the genetic origin of an important fraction of the undiagnosed patients.

In the presented works, we used different sequencing technologies to detect these variants, always validating them with orthogonal approaches. We have also used amplicon-based deep sequencing, PCR and qPCR (TaqMan Copy Number Assay) to validate or better characterize the detected variation. Thus, we propose applying these bioinformatics and experimental approaches to detect these hidden sources of causal genetic variants, which should allow to partially overcome the limitations of the standard MPS-based approaches. Of interest, bioinformatics approaches for the detection of somatic genetic variants can be performed in already generated data. In the clinical practice, routine diagnosis analysis using targeted gene panels are usually performed, and this data often reaches relatively high coverage, which makes them suitable for somatic variant discovery.

Limitations

Some of the main problems of MPS have been previously mentioned. These technologies are not exempt from error and biases. For example, when preparing the libraries for Illumina sequencing, PCR amplification of the genomic material is normally needed, and neutral GC fragments are preferentially amplified (Van Dijk, Jaszczyszyn, and Thermes 2014). Another problem is the error rate of the sequencer itself, and percentages from 0.1% (Fox et al. 2017) to 2% (Kelley, Schatz, and Salzberg 2010) have been reported for Illumina machines. For long read sequencers, this value can be as high as 15% (Rang, Kloosterman, and de Ridder 2018). Once the samples are sequenced, not all the produced reads can be mapped to the reference genome, because there are limitations in read length with respect to the reference (Li and Freudenberg 2014). Besides, there are regions in the genome that are less confidently called. The 1000 Genomes Project has identified the genomic coordinates of such regions after the observation of recurrent errors (Bae et al. 2018).

Apart from all these technical issues, different software for mapping the sequencing reads to a reference genome produce different results (Keel and Snelling 2018; Zhou, Lin, and Xing 2019). The same thing happens when calling germline (Chen et al. 2019) and somatic (Chen et al. 2020) SNVs and indels from short read sequencing, or structural variants with short (Cameron, Di Stefano, and Papenfuss 2019; Whitford et al. 2019) and long read (Luan et al. 2020; Zhou, Lin, and Xing 2019) sequencing methods. Although there are continuous efforts to achieve gold standard methods, like the GATK's best practices, the evolution of the field and the constant incorporation of new technologies complicate this process. To date, it is obvious that the chosen tools will have an impact on the outcome of the study.

I have also highlighted the need to validate the called variants regardless of the technology used and the type of the variant. In this sense, germline SNVs with high quality and sufficient depth (≥ 35X) can reach validation rates of 100% when retrieved from short read sequencing approaches (Zheng et al. 2019). Indels show more problems and, besides, there are also remarkable differences between WGS and WES short read data. In a study that examined this issue, the validation rate for WGS specific indels was 84% in high quality calls, while it dropped to 57% in WES specific indels, even in targeted regions (Fang et al. 2014). In an independent study, the validation rate for SNVs was 96.8%, while for indels it was 82.4% using WGS. In another study, the validation rate for SVs was 97.0% (Werling et al. 2018). Of course, these numbers are always given for variants that have passed through thorough filtering processes and it is not surprising that they are consistent in the bibliography. The reality when working with this type of data is that a lot of promising potential variants are discarded and, more important in medical genomics, that true variants are not called.

Future perspectives

In this thesis, I have used both WES and WGS to discover somatic and germline variants. WES is more cost effective, although the breach between both has been reduced. Besides, WGS is a better strategy for the variant discovery even in coding regions (Belkadi et al. 2015; Meienberg et al. 2016). This fact, along with the possibility of calling non-coding variants and calling SVs more accurately, support a logic transition in the near future to always, or most of the times, perform WGS. But this raises two other issues: much higher coverages are needed for somatic variant calling in bulk tissue, which would increment the price; and considerably more disk space would be required, since the generated files would be larger –and this is not a trivial problem–. This last issue could be solved by using cloud computing methods (Bani Baker et al. 2020) but, again, it will raise the price of the experiments.

The diagnostic rate yielded from MPS greatly varies depending on the type of disease studied and on the type of sequencing –panels, WES or WGS–. In general terms, WGS always obtains the highest diagnostic rates, as has been observed by directly comparing sequencing panels versus WGS (Ellingford et al. 2016) and WES versus WGS (Liu et al. 2019). In a laboratory, 2,509 diagnostic tests were performed from 2012 until 2017 and, in general, their diagnostic rate was 24.1%. This number remained the same over the years, even though the genes they used in their panel increased from 568 to 6,940, which indicates that a small pool of genes is responsible for most of the diagnostics (Hartman et al. 2019). A similar value of 25% was also achieved in 250 probands when using WES (Yang et al. 2013). Apart from this natural shift towards the use of WGS, in upcoming years long read sequencing will allow the discovery of new SVs linked to disease (Mantere, Kersten, and Hoischen 2019; Pollard et al. 2018).

As briefly mentioned above, apart from digging for genes to discover pathogenic variants, some authors are also trying to explore non-coding regions. Some non-coding variants have already been observed to be related to disease, not only in complex diseases but also in Mendelian ones (French and Edwards 2020; Zhang and Lupski 2015). For example, ranking according to predicted regulatory effect on important genes could be a first approach to try to discover new pathogenic non-coding variants (Wells et al. 2019).

Taking into account their potential in discovering SVs and, of course, the fact that coding and non-coding regions can also be explored, a shift toward the use of long read WGS is expected. This would be the natural tendency of MPS as prices are lowered and laboratory and computational methods are refined, especially to reduce their currently high error rate.

Further future perspectives and social considerations

And now, let us imagine. One day we will have the power to sequence the genome of every person on the planet, and we may do it. Besides, we will have every position perfectly covered with no biases or errors. And every chromosome could be sequenced by itself: one read, one whole chromosome (why not?). That way, we will be able to study all types of genetic variation. This idea, considering the evolution of the field in the last two decades, is not crazy. But what will happen if we really sequence the genome of every human being? On the one hand, from the purely scientific and medical point of view, it will be a resource of unquestionable value. But, on the other hand, dangerous ethical concerns will rise regarding privacy rights and the use of this data, something that is already happening today at a lower scale.

In previous years, we have seen how ancestry companies, like 23andMe, are selling data to big pharmaceutical companies with the aim to develop drugs taking into account the genomic profile of the people in their database (Hamzelou 2020). In fact, there are voices that say that these ancestry companies should be paying for their services, and not the other way around, since they earn money by selling their customer's data (Spinney 2020).

Outside the field of genomic data, other companies, with Facebook in the lead, have sold information from their database that served different purposes. Perhaps the most famous example is the Cambridge Analytical scandal, in which this company, by obtaining data from Facebook, built voter profiles and helped the Republican Party campaign to influence the USA elections that they won in 2017 (Confessore 2018). So, the same way that our looks, hobbies and other preferences are information, and information is power, our genomic data is also a valuable source of information. In countries with no universal social security, like the USA, insurance companies dominate the health market. Based on the genomic profile of the individuals, they could adjust the pricing of their services and they could raise it if someone is more likely to develop a disease (Song 2018). This is why the Genetic Information Nondiscrimination Act (GINA) exists in this country. But that is the problem of doing business with health: money becomes more important than lives.

Something similar almost happened involving COVID-19, in which some administrations, like the Community of Madrid, wanted to implement a register of all the COVID-19 tests taken by a person. The idea was to use it for several purposes, like entering in close spaces with more people or even to find a job, but this was stopped due to its discriminatory nature. In the end, it could just be used for health causes by the health care system (Belver 2020).

It is mandatory to implement strong regulatory policies regarding the use of our personal genomic profile information. These rules already exist, especially in clinical practice, but the boundaries are less clear for other types of companies. Only this way discriminatory events could be prevented and stopped.

Concluding remarks

Over the last 20 years, thanks to MPS, our understanding of all fields of knowledge related to genomics has greatly improved, especially in population genetics and medical genomics (Koboldt et al. 2013). These technologies have proven to be a valuable tool for the discovery of a great number of novel genes and pathogenic variants. Thanks to recent advances, like long read sequencing, this number is expected to continue rising. The exploration of other genetic disease models, apart from the monogenic scenario, is also expected to bear more and more promising fruits. In the end, the aim of us scientists is to acquire novel knowledge and to improve the lives of people. And we will keep on trying as long as economical resources and a system focused on making money and being productive allow us to do so.

Conclusions

1. Pathogenic somatic variants in coding regions can be detected from medium/high coverage whole exome sequencing.

2. The analytical approaches to detect somatic variants should consider several variant callers simultaneously. It is also necessary to use several filters to discard false positives, as well as those based on variant annotation to reduce the number of candidate variants. Because of the high number of candidate variants in an individual's exome, the search will probably have to be restricted to a set of candidate genes.

3. Candidate somatic variants must be validated using amplicon-based deep sequencing, which also allows a better estimation of the variant allele frequency. Autoinflammatory diseases (and all primary immunodeficiencies) offer a very good model for somatic variant characterization, given the easy sampling of the tissue of interest and the analysis of different cell populations.

4. The detection of pathogenic structural variants is possible with massive parallel sequencing technologies, with the visual inspection of the mapped reads being particularly useful for undetected variants in candidate genes.

5. Alternative models to single nucleotide variants in coding regions must be considered in the analyses of rare disease patients. The current methodologies already allow to explore this scenario.

List of publications

The following works were also developed during the years of this PhD:

Solís-Moruno, Manuel, Marc de Manuel, Jessica Hernandez-Rodriguez, Claudia Fontsere, Alba Gomara-Castaño, Cristina Valsera-Naranjo, Dietmar Crailsheim, et al. 2017. "Potential Damaging Mutation in LRP5 from Genome Sequencing of the First Reported Chimpanzee with the Chiari Malformation." *Scientific Reports* 7 (1): 15224. https://doi.org/10.1038/s41598-017-15544-w.

Valles-Ibáñez, Guillem de, Ana Esteve-Solé, Mònica Piquer, E. Azucena González-Navarro, Jessica Hernandez-Rodriguez, Hafid Laayouni, Eva González-Roca María Plaza-Martin, Ángela Deyà-Martínez, Andrea Martín-Nalda, Mónica Martínez-Gallo, Marina García-Prat, Lucía del Pino-Molina, Ivón Cuscó, Marta Codina-Solà, Laura Batlle-Masó, **Manuel Solís-Moruno**, Tomàs Marquès-Bonet, Elena Bosch, Eduardo López-Granados, Juan Ignacio Aróstegui, Pere Soler-Palacín, Roger Colobran, Jordi Yagüe, Laia Alsina, Manel Juan,* and Ferran Casals*. 2018. "Evaluating the Genetics of Common Variable Immunodeficiency: Monogenetic Model and Beyond." *Frontiers in Immunology* 9 (MAY): 1–15. https://doi.org/10.3389/fimmu.2018.00636.

Batlle-Masó, Laura, Anna Mensa-Vilaró, **Manuel Solís-Moruno**, Tomàs Marquès-Bonet, Juan I. Arostegui, and Ferran Casals. 2020. "Genetic Diagnosis of Autoinflammatory Disease Patients Using Clinical Exome Sequencing." *European Journal of Medical Genetics* 63 (5): 103920. https://doi.org/10.1016/j.ejmg.2020.103920.

Lobon, Irene, **Manuel Solís-Moruno**, David Juan, Ashraf Muhaisen, Federico Abascal, Paula Esteller-Cucala, Raquel García-Pérez, et al. 2020. "Somatic Mutations in Parkinson Disease Are Enriched in Synaptic and Neuronal Processes." *MedRxiv*. https://doi.org/10.1101/2020.09.14.20190538

Solís-Moruno, **Manuel,** Juan I. Aróstegui, and Ferran Casals. "Somatic genetic variation in healthy tissue and non-cancer diseases." Manuscript in preparation.

Bibliography

Abascal, Federico, Luke M R Harvey, Emily Mitchell, Andrew R J Lawson, Stefanie V Lensing, Peter Ellis, Andrew J C Russell, et al. 2021. "Somatic Mutation Landscapes at Single-Molecule Resolution." *Nature* 593 (7859): 405–10. https://doi.org/10.1038/s41586-021-03477-4.

Abelson, Sagi, Grace Collord, Stanley W.K. Ng, Omer Weissbrod, Netta Mendelson Cohen, Elisabeth Niemeyer, Noam Barda, et al. 2018. "Prediction of Acute Myeloid Leukaemia Risk in Healthy Individuals." *Nature* 559 (7714): 400–404. https://doi.org/10.1038/s41586-018-0317-6.

Abyzov, Alexej, Alexander E. Urban, Michael Snyder, and Mark Gerstein. 2011. "CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing." *Genome Research* 21 (6): 974–84. https://doi.org/10.1101/gr.114876.110.

Adzhubei, Ivan A., Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. "A Method and Server for Predicting Damaging Missense Mutations." *Nature Methods* 7 (4): 248–49. https://doi.org/10.1038/nmeth0410-248.

Aksentijevich, Ivona, Seth L. Masters, Polly J. Ferguson, Paul Dancey, Joost Frenkel, Annet van Royen-Kerkhoff, Ron Laxer, et al. 2009. "An Autoinflammatory Disease with Deficiency of the Interleukin-1–Receptor Antagonist." *New England Journal of Medicine* 360 (23): 2426–37. https://doi.org/10.1056/nejmoa0807865.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. "The Repertoire of Mutational Signatures in Human Cancer." *Nature* 578 (7793): 94–101. https://doi.org/10.1038/s41586-020-1943-3.

Alexandrov, Ludmil B., and Michael R. Stratton. 2014. "Mutational Signatures: The Patterns of Somatic Mutations Hidden in Cancer Genomes." *Current Opinion in Genetics and Development* 24 (1): 52–60. https://doi.org/10.1016/j.gde.2013.11.014.

Alkan, Can, Bradley P. Coe, and Evan E. Eichler. 2011. "Genome Structural Variation Discovery and Genotyping." *Nature Reviews Genetics* 12 (5): 363–76. https://doi.org/10.1038/nrg2958.

Alkan, Can, Jeffrey M. Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O. Kitzman, et al. 2009. "Personalized Copy Number and Segmental Duplication Maps Using Next-Generation Sequencing." *Nature Genetics* 41 (10): 1061–67. https://doi.org/10.1038/ng.437.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Anne Marie E. Welch, Max L. Dougherty, et al. 2019. "Characterizing the Major Structural Variant Alleles of the Human Genome." *Cell* 176 (3): 663-675.e19. https://doi.org/10.1016/j.cell.2018.12.019.

Badano, Jose L., and Nicholas Katsanis. 2002. "Beyond Mendel: An Evolving View of Human Genetic Disease Transmission." *Nature Reviews Genetics* 3 (10): 779–89. https://doi.org/10.1038/nrg910.

Bae, Taejeong, Livia Tomasini, Jessica Mariani, Bo Zhou, Tanmoy Roychowdhury, Daniel Franjic, Mihovil Pletikos, et al. 2018. "Different Mutational Rates and Mechanisms in Human Cells at Pregastrulation and Neurogenesis." *Science* 359 (6375): 550–55. https://doi.org/10.1126/science.aan8690.

Bani Baker, Qanita, Mahmoud Hammad, Wesam Al-Rashdan, Yaser Jararweh, Mohammad AL-Smadi, and Mohammad Al-Zinati. 2020. "Comprehensive

Comparison of Cloud-Based NGS Data Analysis and Alignment Tools." *Informatics in Medicine Unlocked* 18: 100296. https://doi.org/10.1016/j.imu.2020.100296.

Barker, Nick, Meritxell Huch, Pekka Kujala, Marc van de Wetering, Hugo J. Snippert, Johan H. van Es, Toshiro Sato, et al. 2010. "Lgr5+ve Stem Cells Drive Self-Renewal in the Stomach and Build Long-Lived Gastric Units In Vitro." *Cell Stem Cell* 6 (1): 25–36. https://doi.org/10.1016/j.stem.2009.11.013.

Bartl, Simona, Meredith Baish, Irving L. Weissman, and Marilyn Diaz. 2003. "Did the Molecules of Adaptive Immunity Evolve from the Innate Immune System?" *Integrative and Comparative Biology* 43 (2): 338–46. https://doi.org/10.1093/icb/43.2.338.

Belkadi, Aziz, Alexandre Bolze, Yuval Itan, Aurélie Cobat, Quentin B. Vincent, Alexander Antipenko, Lei Shang, Bertrand Boisson, Jean Laurent Casanova, and Laurent Abel. 2015. "Whole-Genome Sequencing Is More Powerful than Whole-Exome Sequencing for Detecting Exome Variants." *Proceedings of the National Academy of Sciences of the United States of America* 112 (17): 5473–78. https://doi.org/10.1073/pnas.1418631112.

Belver, Marta. 2020. "La Cartilla Covid de Ayuso Empezará a Funcionar En Madrid El Próximo Lunes." *El Mundo*, 2020. https://www.elmundo.es/madrid/2020/12/10/5fd1eb3ffc6c8365038b4722.html.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456 (7218): 53–59. https://doi.org/10.1038/nature07517.

Blokzijl, Francis, Joep De Ligt, Myrthe Jager, Valentina Sasselli, Sophie Roerink, Nobuo Sasaki, Meritxell Huch, et al. 2016. "Tissue-Specific Mutation Accumulation in Human Adult Stem Cells during Life." *Nature* 538 (7624): 260–64. https://doi.org/10.1038/nature19768.

Brehm, Anja, Yin Liu, Afzal Sheikh, Bernadette Marrero, Ebun Omoyinmi, Qing Zhou, Gina Montealegre, et al. 2015. "Additive Loss-of-Function Proteasome Subunit Mutations in CANDLE/PRAAS Patients Promote Type i IFN Production." *Journal of Clinical Investigation* 125 (11): 4196–4211. https://doi.org/10.1172/JCI81260.

Brunner, Simon F, Nicola D Roberts, Luke A Wylie, Luiza Moore, Sarah J Aitken, Susan E Davies, Mathijs A Sanders, et al. 2019. "Somatic Mutations and Clonal Dynamics in Healthy and Cirrhotic Human Liver." *Nature* 574 (November 2018).

Burnet, F. M. 1976. "A Modification of Jerne's Theory of Antibody Production Using the Concept of Clonal Selection." *CA: A Cancer Journal for Clinicians* 26 (2): 119–21. https://doi.org/10.3322/canjclin.26.2.119.

Cai, Lei, Wei Yuan, Zhou Zhang, Lin He, and Kuo Chen Chou. 2016. "In-Depth Comparison of Somatic Point Mutation Callers Based on Different Tumor next-Generation Sequencing Depth Data." *Scientific Reports* 6 (November): 1–9. https://doi.org/10.1038/srep36540.

Callari, Maurizio, Stephen John Sammut, Leticia De Mattos-Arruda, Alejandra Bruna, Oscar M. Rueda, Suet Feung Chin, and Carlos Caldas. 2017. "Intersect-Then-Combine Approach: Improving the Performance of Somatic Variant Calling in Whole Exome Sequencing Data Using Multiple Aligners and Callers." *Genome Medicine* 9 (1): 1–11. https://doi.org/10.1186/s13073-017-0425-1.

Cameron, Daniel L., Leon Di Stefano, and Anthony T. Papenfuss. 2019. "Comprehensive Evaluation and Characterisation of Short Read General-Purpose Structural Variant Calling Software." *Nature Communications* 10 (1): 1–11. https://doi.org/10.1038/s41467-019-11146-4.

Campbell, Peter J., Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, et al. 2020. "Pan-Cancer Analysis of Whole Genomes." *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Chen, Jiayun, Xingsong Li, Hongbin Zhong, Yuhuan Meng, and Hongli Du. 2019. "Systematic Comparison of Germline Variant Calling Pipelines Cross Multiple Next-Generation Sequencers." *Scientific Reports* 9 (1): 1–13. https://doi.org/10.1038/s41598-019-45835-3.

Chen, Zixi, Yuchen Yuan, Xiaoshi Chen, Jiayun Chen, Shudai Lin, Xingsong Li, and Hongli Du. 2020. "Systematic Comparison of Somatic Variant Calling Performance among Different Sequencing Depth and Mutation Frequency." *Scientific Reports* 10 (1): 1–9. https://doi.org/10.1038/s41598-020-60559-5.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff." *Fly* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

Collins, Francis S., Eric D. Green, Alan E. Guttmacher, and Mark S. Guyer. 2003. "A Vision for the Future of Genomics Research." *Nature* 422 (6934): 835–47. https://doi.org/10.1038/nature01626.

Condit, Celeste M., Paul J. Achter, Ilon Lauer, and Enid Sefcovic. 2002. "The Changing Meanings of 'Mutation:' A Contextualized Study of Public Discourse." *Human Mutation* 19 (1): 69–75. https://doi.org/10.1002/humu.10023.

Confessore, Nicholas. 2018. "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far." *The New York Times*, 2018. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

Cooper, Max D., and Matthew N. Alder. 2006. "The Evolution of Adaptive Immune Systems." *Cell* 124 (4): 815–22. https://doi.org/10.1016/j.cell.2006.02.001.

Costain, Gregory, Rebekah Jobling, Susan Walker, Miriam S. Reuter, Meaghan Snell, Sarah Bowdin, Ronald D. Cohn, et al. 2018. "Periodic Reanalysis of Whole-Genome Sequencing Data Enhances the Diagnostic Advantage over Standard Clinical Genetic Testing." *European Journal of Human Genetics* 26 (5): 740–44. https://doi.org/10.1038/s41431-018-0114-6.

Cuadrado, Eloy, Adeline Vanderver, Kristy J. Brown, Annie Sandza, Asako Takanohashi, Machiel H. Jansen, Jasper Anink, et al. 2015. "Aicardi-Goutières Syndrome Harbours Abundant Systemic and Brain-Reactive Autoantibodies." *Annals of the Rheumatic Diseases* 74 (10): 1931–39. https://doi.org/10.1136/annrheumdis-2014-205396.

Cunningham, John A., Alexander G. Liu, Stefan Bengtson, and Philip C.J. Donoghue. 2017. "The Origin of Animals: Can Molecular Clocks and the Fossil Record Be Reconciled?" *BioEssays* 39 (1): 1–12. https://doi.org/10.1002/bies.201600120.

D'Osualdo, Andrea, Paolo Picco, Francesco Caroli, Marco Gattorno, Raffaella Giacchino, Patrizia Fortini, Fabrizia Corona, et al. 2005. "MVK Mutations and Associated Clinical Features in Italian Patients Affected with Autoinflammatory Disorders and Recurrent Fever." *European Journal of Human Genetics* 13 (3): 314–

20. https://doi.org/10.1038/sj.ejhg.5201323.

Davis, Brigid M., Michael C. Chao, and Matthew K. Waldor. 2013. "Entering the Era of Bacterial Epigenomics with Single Molecule Real Time DNA Sequencing." *Current Opinion in Microbiology* 16 (2): 192–98. https://doi.org/10.1016/j.mib.2013.01.011.

Dayer, Jean Michel, Francesca Oliviero, and Leonardo Punzi. 2017. "A Brief History of IL-1 and IL-1 Ra in Rheumatology." *Frontiers in Pharmacology* 8 (MAY): 1–8. https://doi.org/10.3389/fphar.2017.00293.

Desai, Pinkal, Nuria Mencia-Trinchant, Oleksandr Savenkov, Michael S. Simon, Gloria Cheang, Sangmin Lee, Michael Samuel, et al. 2018. "Somatic Mutations Precede Acute Myeloid Leukemia Years before Diagnosis." *Nature Medicine* 24 (7): 1015–23. https://doi.org/10.1038/s41591-018-0081-z.

Dijk, Erwin L. Van, Yan Jaszczyszyn, and Claude Thermes. 2014. "Library Preparation Methods for Next-Generation Sequencing: Tone down the Bias." *Experimental Cell Research* 322 (1): 12–20. https://doi.org/10.1016/j.yexcr.2014.01.008.

Doria, A., M. Zen, S. Bettio, M. Gatto, N. Bassi, L. Nalotto, A. Ghirardello, L. Iaccarino, and L. Punzi. 2012. "Autoinflammation and Autoimmunity: Bridging the Divide." *Autoimmunity Reviews* 12 (1): 22–30. https://doi.org/10.1016/j.autrev.2012.07.018.

Edge, Peter, and Vikas Bansal. 2019. "Longshot Enables Accurate Variant Calling in Diploid Genomes from Single-Molecule Long Read Sequencing." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-12493-y.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. "Real-Time DNA Sequencing from Single Polymerase Molecules." *Science* 323 (5910): 133–38. https://doi.org/10.1126/science.1162986.

Eisen, J. A., J. F. Heidelberg, O. White, and S. L. Salzberg. 2000. "Evidence for Symmetric Chromosomal Inversions around the Replication Origin in Bacteria." *Genome Biology* 1 (6): 1–9. https://doi.org/10.1186/gb-2000-1-6-research0011.

Ellingford, Jamie M., Stephanie Barton, Sanjeev Bhaskar, Simon G. Williams, Panagiotis I. Sergouniotis, James O'Sullivan, Janine A. Lamb, et al. 2016. "Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease." *Ophthalmology* 123 (5): 1143–50. https://doi.org/10.1016/j.ophtha.2016.01.009.

Fang, Han, Yiyang Wu, Giuseppe Narzisi, Jason A ORawe, Laura T Jimenez Barrón, Julie Rosenbaum, Michael Ronemus, Ivan Iossifov, Michael C. Schatz, and Gholson J. Lyon. 2014. "Reducing INDEL Calling Errors in Whole Genome and Exome Sequencing Data." *Genome Medicine* 6 (10): 89. https://doi.org/10.1186/s13073-014-0089-z.

Feng, Zhixing, Gang Fang, Jonas Korlach, Tyson Clark, Khai Luong, Xuegong Zhang, Wing Wong, and Eric Schadt. 2013. "Detecting DNA Modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic." *PLoS Computational Biology* 9 (3): 1–10. https://doi.org/10.1371/journal.pcbi.1002935.

Feuk, Lars, Andrew R. Carson, and Stephen W. Scherer. 2006. "Structural Variation in the Human Genome." *Nature Reviews Genetics* 7 (2): 85–97. https://doi.org/10.1038/nrg1767.

Fisher, Elizabeth, and Peter Scambler. 1994. "Human Haploinsufficiency — One for Sorrow, Two for Joy." *Nature Genetics* 7 (1): 5–7.

https://doi.org/10.1038/ng0594-5.

Flajnik, Martin F., and Masanori Kasahara. 2010. "Origin and Evolution of the Adaptive Immune System: Genetic Events and Selective Pressures." *Nature Reviews Genetics* 11 (1): 47–59. https://doi.org/10.1038/nrg2703.

Fox, Edward J, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb. 2017. "Accuracy of Next Generation Sequencing Platforms." *Next Generation, Sequencing & Applications* 1 (03): 1–5. https://doi.org/10.4172/jngsa.1000106.

Frazer, Kelly A., Sarah S. Murray, Nicholas J. Schork, and Eric J. Topol. 2009. "Human Genetic Variation and Its Contribution to Complex Traits." *Nature Reviews Genetics* 10 (4): 241–51. https://doi.org/10.1038/nrg2554.

French, J. D., and S. L. Edwards. 2020. "The Role of Noncoding Variants in Heritable Disease." *Trends in Genetics* 36 (11): 880–91. https://doi.org/10.1016/j.tig.2020.07.004.

Fromer, Menachem, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker, et al. 2012. "Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth." *American Journal of Human Genetics* 91 (4): 597–607. https://doi.org/10.1016/j.ajhg.2012.08.005.

Gazzo, Andrea, Daniele Raimondi, Dorien Daneels, Yves Moreau, Guillaume Smits, Sonia Van Dooren, and Tom Lenaerts. 2017. "Understanding Mutational Effects in Digenic Diseases." *Nucleic Acids Research* 45 (15): 1–11. https://doi.org/10.1093/nar/gkx557.

Gellert, Martin. 1992. "Molecular Analysis of V(D)J Recombination." *Annual Review of Genetics* 26 (1): 425–46. https://doi.org/10.1146/annurev.ge.26.120192.002233.

Genovese, Giulio, Anna K. Kähler, Robert E. Handsaker, Johan Lindberg, Samuel A. Rose, Samuel F. Bakhoum, Kimberly Chambert, et al. 2014. "Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence." *New England Journal of Medicine* 371 (26): 2477–87. https://doi.org/10.1056/NEJMoa1409405.

Georgin-Lavialle, Sophie, Stéphanie Ducharme-Benard, Guillaume Sarrabay, Léa Savey, Gilles Grateau, and Véronique Hentgen. 2020. "Systemic Autoinflammatory Diseases: Clinical State of the Art." *Best Practice and Research: Clinical Rheumatology* 34 (4). https://doi.org/10.1016/j.berh.2020.101529.

Gery, Igal, Richard K. Gershon, and Byron H. Waksman. 1972. "Potentiation of the T-Lvmphocyte Response to Mitogens: I. The Responding Cell." *Journal of Experimental Medicine* 136 (1): 128–42. https://doi.org/10.1084/jem.136.1.128.

Gilissen, Christian, Alexander Hoischen, Han G. Brunner, and Joris A. Veltman. 2011. "Unlocking Mendelian Disease Using Exome Sequencing." *Genome Biology* 12 (9). https://doi.org/10.1186/gb-2011-12-9-228.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews Genetics* 17 (6): 333–51. https://doi.org/10.1038/nrg.2016.49.

Gouil, Quentin, and Andrew Keniry. 2019. "Latest Techniques to Study DNA Methylation." Edited by Marnie Blewitt. *Essays in Biochemistry* 63 (6): 639–48. https://doi.org/10.1042/EBC20190027.

Graham, Daniel B., and Ramnik J. Xavier. 2020. "Pathway Paradigms Revealed from the Genetics of Inflammatory Bowel Disease." *Nature* 578 (7796): 527–39. https://doi.org/10.1038/s41586-020-2025-2.

Gray, J. W., A. V. Carrano, L. L. Steinmetz, M. A. Van Dilla, D. H. Moore IInd, B. H.

Mayall, and M. L. Mendelsohn. 1975. "Chromosome Measurement and Sorting by Flow Systems." *Proceedings of the National Academy of Sciences of the United States of America* 72 (4): 1231–34. https://doi.org/10.1073/pnas.72.4.1231.

Gritsenko, Anna, Shi Yu, Fatima Martin-Sanchez, Ines Diaz-del-Olmo, Eva-Maria Nichols, Daniel M. Davis, David Brough, and Gloria Lopez-Castejon. 2020. "Priming Is Dispensable for NLRP3 Inflammasome Activation in Human Monocytes In Vitro." *Frontiers in Immunology* 11 (September): 1–14. https://doi.org/10.3389/fimmu.2020.565924.

Gudbjartsson, Daniel F., Hannes Helgason, Sigurjon A. Gudjonsson, Florian Zink, Asmundur Oddson, Arnaldur Gylfason, Soren Besenbacher, et al. 2015. "Large-Scale Whole-Genome Sequencing of the Icelandic Population." *Nature Genetics* 47 (5): 435–44. https://doi.org/10.1038/ng.3247.

Guo, Haitao, Justin B. Callaway, and Jenny P.Y. Ting. 2015. "Inflammasomes: Mechanism of Action, Role in Disease, and Therapeutics." *Nature Medicine* 21 (7): 677–87. https://doi.org/10.1038/nm.3893.

Hach, Faraz, Iman Sarrafi, Farhad Hormozdiari, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. 2014. "MrsFAST-Ultra: A Compact, SNP-Aware Mapper for High Performance Sequencing Applications." *Nucleic Acids Research* 42 (W1): 494–500. https://doi.org/10.1093/nar/gku370.

Hafner-Bratkovič, Iva, Petra Sušjan, Duško Lainšček, Ana Tapia-Abellán, Kosta Cerović, Lucija Kadunc, Diego Angosto-Bazarra, Pablo Pelegrin, and Roman Jerala. 2018. "NLRP3 Lacking the Leucine-Rich Repeat Domain Can Be Fully Activated via the Canonical Inflammasome Pathway." *Nature Communications* 9 (1): 5182. https://doi.org/10.1038/s41467-018-07573-4.

Hamzelou, Jessica. 2020. "23andMe Has Sold the Rights to Develop a Drug Based on Its Users' DNA." *New Scientist*, 2020. https://www.newscientist.com/article/2229828-23andme-has-sold-the-rights-to-develop-a-drug-based-on-its-users-dna/.

Hartman, Paige, Kenneth Beckman, Kevin Silverstein, Sophia Yohe, Matthew Schomaker, Christine Henzler, Getiria Onsongo, et al. 2019. "Next Generation Sequencing for Clinical Diagnostics : Five Year Experience of an Academic Laboratory." *Molecular Genetics and Metabolism Reports* 19 (February): 100464. https://doi.org/10.1016/j.ymgmr.2019.100464.

Heather, James M., and Benjamin Chain. 2016. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics* 107 (1): 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.

Heller, David, and Martin Vingron. 2019. "SVIM: Structural Variant Identification Using Mapped Long Reads." *Bioinformatics* 35 (17): 2907–15. https://doi.org/10.1093/bioinformatics/btz041.

———. 2020. "SVIM-Asm: Structural Variant Detection from Haploid and Diploid Genome Assemblies." *Bioinformatics*, no. December: 1–3. https://doi.org/10.1093/bioinformatics/btaa1034.

Hindorff, Lucia A., Vence L. Bonham, Lawrence C. Brody, Margaret E.C. Ginoza, Carolyn M. Hutter, Teri A. Manolio, and Eric D. Green. 2018. "Prioritizing Diversity in Human Genomics Research." *Nature Reviews Genetics* 19 (3): 175–85. https://doi.org/10.1038/nrg.2017.89.

Hoffman, Hal M., and Lori Broderick. 2017. "Editorial: It Just Takes One: Somatic Mosaicism in Autoinflammatory Disease." *Arthritis & Rheumatology* 69 (2): 253–56. https://doi.org/10.1002/art.39961.

Hoffman, Hal M., James L. Mueller, David H. Broide, Alan A. Wanderer, and Richard

D. Kolodner. 2001. "Mutation of a New Gene Encoding a Putative Pyrin-like Protein Causes Familial Cold Autoinflammatory Syndrome and Muckle-Wells Syndrome." *Nature Genetics* 29 (3): 301–5. https://doi.org/10.1038/ng756.

Hofmann, Ariane L., Jonas Behr, Jochen Singer, Jack Kuipers, Christian Beisel, Peter Schraml, Holger Moch, and Niko Beerenwinkel. 2017. "Detailed Simulation of Cancer Exome Sequencing Data Reveals Differences and Common Limitations of Variant Callers." *BMC Bioinformatics* 18 (1): 1–15. https://doi.org/10.1186/s12859-016-1417-7.

Holley, Guillaume, Doruk Beyter, Helga Ingimundardottir, Peter L. Møller, Snædis Kristmundsdottir, Hannes P. Eggertsson, and Bjarni V. Halldorsson. 2021. "Ratatosk: Hybrid Error Correction of Long Reads Enables Accurate Variant Calling and Assembly." *Genome Biology* 22 (1): 1–22. https://doi.org/10.1186/s13059-020-02244-4.

Horai, Reiko, Masahide Asano, Katsuko Sudo, Hirotaka Kanuka, Masatoshi Suzuki, Masugi Nishihara, Michio Takahashi, and Yoichiro Iwakura. 1998. "Production of Mice Deficient in Genes for Interleukin (IL)-1α, IL-1β, IL-1α/β, and IL-1 Receptor Antagonist Shows That IL-1β Is Crucial in Turpentine-Induced Fever Development and Glucocorticoid Secretion." *Journal of Experimental Medicine* 187 (9): 1463–75. https://doi.org/10.1084/jem.187.9.1463.

Horebeek, L. Van, B. Dubois, and A. Goris. 2019. "Somatic Variants: New Kids on the Block in Human Immunogenetics." *Trends in Genetics* 35 (12): 935–47. https://doi.org/10.1016/j.tig.2019.09.005.

Inocencio, Jaime De, Anna Mensa-Vilaro, Pilar Tejada-Palacios, Eugenia Enriquez-Merayo, Eva González-Roca, Giuliana Magri, Estibaliz Ruiz-Ortiz, Andrea Cerutti, Jordi Yagüe, and Juan I. Aróstegui. 2015. "Somatic NOD2 Mosaicism in Blau Syndrome." *Journal of Allergy and Clinical Immunology* 136 (2): 484-487.e2. https://doi.org/10.1016/j.jaci.2014.12.1941.

International HapMap Consortium. 2003. "The International HapMap Project." *Nature* 426 (6968): 789–96. https://doi.org/10.1038/nature02168.

International Human Genome Sequencing Consortium. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

———. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431 (7011): 931–45.

Jaiswal, Siddhartha, and Benjamin L. Ebert. 2019. "Clonal Hematopoiesis in Human Aging and Disease." *Science* 366 (6465). https://doi.org/10.1126/science.aan4673.

Jaiswal, Siddhartha, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V. Grauman, Brenton G. Mar, R. Coleman Lindsley, et al. 2014. "Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes." *New England Journal of Medicine* 371 (26): 2488–98. https://doi.org/10.1056/NEJMoa1408617.

Janeway, C. A. 1989. "Approaching the Asymptote? Evolution and Revolution in Immunology." *Cold Spring Harbor Symposia on Quantitative Biology* 54 (1): 1–13. https://doi.org/10.1101/sqb.1989.054.01.003.

Jarne, Philippe, and Pierre J.L. Lagoda. 1996. "Microsatellites, from Molecules to Populations and Back." *Trends in Ecology and Evolution* 11 (10): 424–29. https://doi.org/10.1016/0169-5347(96)10049-5.

Jones, David, Keiran M. Raine, Helen Davies, Patrick S. Tarpey, Adam P. Butler, Jon W. Teague, Serena Nik-Zainal, and Peter J. Campbell. 2016.

"CgpCaVEManWrapper: Simple Execution of Caveman in Order to Detect Somatic Single Nucleotide Variants in NGS Data." *Current Protocols in Bioinformatics* 2016 (December): 15.10.1-15.10.18. https://doi.org/10.1002/cpbi.20.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581 (7809): 434–43. https://doi.org/10.1038/s41586-020-2308-7.

Kaufmann, Stefan H E. 2008. "Immunology's Foundation: The 100-Year Anniversary of the Nobel Prize to Paul Ehrlich and Elie Metchnikoff." *Nature Immunology* 9 (7): 705–12.

Kawasaki, Yuri, Hirotsugu Oda, Jun Ito, Akira Niwa, Takayuki Tanaka, Atsushi Hijikata, Ryosuke Seki, et al. 2017. "Identification of a High-Frequency Somatic NLRC4 Mutation as a Cause of Autoinflammation by Pluripotent Cell–Based Phenotype Dissection." *Arthritis and Rheumatology* 69 (2): 447–59. https://doi.org/10.1002/art.39960.

Kchouk, Mehdi, Jean Francois Gibrat, and Mourad Elloumi. 2017. "Generations of Sequencing Technologies: From First to Next Generation." *Biology and Medicine* 09 (03). https://doi.org/10.4172/0974-8369.1000395.

Keel, Brittney N., and Warren M. Snelling. 2018. "Comparison of Burrows-Wheeler Transform-Based Mapping Algorithms Used in High-Throughput Whole-Genome Sequencing: Application to Illumina Data for Livestock Genomes 1." *Frontiers in Genetics* 9 (FEB): 1–6. https://doi.org/10.3389/fgene.2018.00035.

Kelley, David R., Michael C. Schatz, and Steven L. Salzberg. 2010. "Quake: Quality-Aware Detection and Correction of Sequencing Errors." *Genome Biology* 11 (11). https://doi.org/10.1186/gb-2010-11-11-r116.

Kim, Sangtae, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, et al. 2018. "Strelka2: Fast and Accurate Calling of Germline and Somatic Variants." *Nature Methods* 15 (8): 591–94. https://doi.org/10.1038/s41592-018-0051-x.

Kim, Su Y., Laurent Jacob, and Terence P. Speed. 2014. "Combining Calls from Multiple Somatic Mutation-Callers." *BMC Bioinformatics* 15 (1): 1–8. https://doi.org/10.1186/1471-2105-15-154.

King, David G., Morris Soller, and Yechezkel Kashi. 1997. "Evolutionary Tuning Knobs." *Endeavour* 21 (1): 36–40. https://doi.org/10.1016/S0160-9327(97)01005-3.

Koboldt, Daniel C., Karyn Meltz Steinberg, David E. Larson, Richard K. Wilson, and Elaine R. Mardis. 2013. "The Next-Generation Sequencing Revolution and Its Impact on Genomics." *Cell* 155 (1): 27–38. https://doi.org/10.1016/j.cell.2013.09.006.

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." *Genome Research* 22 (3): 568–76. https://doi.org/10.1101/gr.129684.111.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive *κ*-Mer Weighting and Repeat Separation." *Genome Research* 27 (5): 722–36. https://doi.org/10.1101/gr.215087.116.

Kousi, Maria, and Nicholas Katsanis. 2015. "Genetic Modifiers and Oligogenic

Inheritance." *Cold Spring Harbor Perspectives in Medicine* 5 (6): 1–22. https://doi.org/10.1101/cshperspect.a017145.

Krawczak, Michael, and David N. Cooper. 1991. "Gene Deletions Causing Human Genetic Disease: Mechanisms of Mutagenesis and the Role of the Local DNA Sequence Environment." *Human Genetics* 86 (5): 425–41. https://doi.org/10.1007/BF00194629.

Krøigård, Anne Bruun, Mads Thomassen, Anne Vibeke Lænkholm, Torben A. Kruse, and Martin Jakob Larsen. 2016. "Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data." *PLoS ONE* 11 (3): 1–15. https://doi.org/10.1371/journal.pone.0151664.

Krumm, Niklas, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, Aaron R. Quinlan, Deborah A. Nickerson, and Evan E. Eichler. 2012. "Copy Number Variation Detection and Genotyping from Exome Sequence Data." *Genome Research* 22 (8): 1525–32. https://doi.org/10.1101/gr.138115.112.

Kuderna, Lukas F.K., Esther Lizano, Eva Julià, Jessica Gomez-Garrido, Aitor Serres-Armero, Martin Kuhlwilm, Regina Antoni Alandes, et al. 2019. "Selective Single Molecule Sequencing and Assembly of a Human Y Chromosome of African Origin." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-018-07885-5.

Kumar, Kishore R., Mark J. Cowley, and Ryan L. Davis. 2019. "Next-Generation Sequencing and Emerging Technologies." *Seminars in Thrombosis and Hemostasis* 45 (7): 661–73. https://doi.org/10.1055/s-0039-1688446.

Lachmann, Helen J., Isabelle Kone-Paut, Jasmin B. Kuemmerle-Deschner, Kieron S. Leslie, Eric Hachulla, Pierre Quartier, Xavier Gitton, Albert Widmer, Neha Patel, and Philip N. Hawkins. 2009. "Use of Canakinumab in the Cryopyrin-Associated Periodic Syndrome." *New England Journal of Medicine* 360 (23): 2416–25. https://doi.org/10.1056/nejmoa0810787.

Lai, Zhongwu, Aleksandra Markovets, Miika Ahdesmaki, Brad Chapman, Oliver Hofmann, Robert Mcewen, Justin Johnson, Brian Dougherty, J. Carl Barrett, and Jonathan R. Dry. 2016. "VarDict: A Novel and Versatile Variant Caller for next-Generation Sequencing in Cancer Research." *Nucleic Acids Research* 44 (11): 1–11. https://doi.org/10.1093/nar/gkw227.

Lanier, Lewis L., and Joseph C. Sun. 2009. "Do the Terms Innate and Adaptive Immunity Create Conceptual Barriers?" *Nature Reviews Immunology* 9 (5): 302–3. https://doi.org/10.1038/nri2547.

Lawson, Andrew R J, Federico Abascal, Tim H H Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, et al. 2020. "Extensive Heterogeneity in Somatic Mutation and Selection in the Human Bladder." *Science* 370 (6512): 75–82. https://doi.org/10.1126/science.aba8347.

Lee-Six, Henry, Sigurgeir Olafsson, Peter Ellis, Robert J. Osborne, Mathijs A. Sanders, Luiza Moore, Nikitas Georgakopoulos, et al. 2019. "The Landscape of Somatic Mutation in Normal Colorectal Epithelial Cells." *Nature* 574 (7779): 532–37. https://doi.org/10.1038/s41586-019-1672-7.

Levy, Samuel, Granger Sutton, Pauline C. Ng, Lars Feuk, Aaron L. Halpern, Brian P. Walenz, Nelson Axelrod, et al. 2007. "The Diploid Genome Sequence of an Individual Human." *PLoS Biology* 5 (10): 2113–44. https://doi.org/10.1371/journal.pbio.0050254.

Li, Chun, and Scott M. Williams. 2013. "Human Somatic Variation: It's Not Just for

Cancer Anymore." *Current Genetic Medicine Reports* 1 (4): 212–18. https://doi.org/10.1007/s40142-013-0029-z.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *ArXiv:1303.3997 [q-Bio.GN]*, March. http://arxiv.org/abs/1303.3997.

———. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Wentian, and Jan Freudenberg. 2014. "Mappability and Read Length." *Frontiers in Genetics* 5 (NOV): 1–1. https://doi.org/10.3389/fgene.2014.00381.

Liu, Hong-Yan, Liyuan Zhou, Meng-Yue Zheng, Jia Huang, Shu Wan, Aiying Zhu, Mingjie Zhang, et al. 2019. "Diagnostic and Clinical Utility of Whole Genome Sequencing in a Cohort of Undiagnosed Chinese Families with Rare Diseases." *Scientific Reports* 9 (1): 19365. https://doi.org/10.1038/s41598-019-55832-1.

Liu, Yin, Adriana A. Jesus, Bernadette Marrero, Dan Yang, Suzanne E. Ramsey, Gina A. Montealegre Sanchez, Klaus Tenbrock, et al. 2014. "Activated STING in a Vascular and Pulmonary Syndrome." *New England Journal of Medicine* 371 (6): 507–18. https://doi.org/10.1056/NEJMoa1312625.

Lobon, Irene, Manuel Solís-Moruno, David Juan, Ashraf Muhaisen, Federico Abascal, Paula Esteller-Cucala, Raquel García-Pérez, et al. 2020. "Somatic Mutations in Parkinson Disease Are Enriched in Synaptic and Neuronal Processes." *MedRxiv*, 2020.09.14.20190538. https://doi.org/10.1101/2020.09.14.20190538.

Lodato, Michael A., Rachel E. Rodin, Craig L. Bohrson, Michael E. Coulter, Alison R. Barton, Minseok Kwon, Maxwell A. Sherman, et al. 2018. "Aging and Neurodegeneration Are Associated with Increased Mutations in Single Human Neurons." *Science* 359 (6375): 555–59. https://doi.org/10.1126/science.aao4426.

Lodato, Michael A., Mollie B. Woodworth, Semin Lee, Gilad D. Evrony, Bhaven K. Mehta, Amir Karger, Soohyun Lee, et al. 2015. "Somatic Mutation in Single Human Neurons Tracks Developmental and Transcriptional History." *Science* 350 (6256): 94–98. https://doi.org/10.1126/science.aab1785.

Loddo, Italia, and Claudio Romano. 2015. "Inflammatory Bowel Disease: Genetics, Epigenetics, and Pathogenesis." *Frontiers in Immunology* 6 (NOV): 6–11. https://doi.org/10.3389/fimmu.2015.00551.

Logsdon, Glennis A., Mitchell R. Vollger, and Evan E. Eichler. 2020. "Long-Read Human Genome Sequencing and Its Applications." *Nature Reviews Genetics* 21 (10): 597–614. https://doi.org/10.1038/s41576-020-0236-x.

Lopez-Garcia, Carlos, Allon M. Klein, Benjamin D. Simons, and Douglas J. Winton. 2010. "Intestinal Stem Cell Replacement Follows a Pattern of Neutral Drift." *Science* 330 (6005): 822–25. https://doi.org/10.1126/science.1196236.

Lu, Hengyun, Francesca Giordano, and Zemin Ning. 2016. "Oxford Nanopore MinION Sequencing and Genome Assembly." *Genomics, Proteomics and Bioinformatics* 14 (5): 265–79. https://doi.org/10.1016/j.gpb.2016.05.004.

Luan, Mei Wei, Xiao Ming Zhang, Zi Bin Zhu, Ying Chen, and Shang Qian Xie. 2020. "Evaluating Structural Variation Detection Tools for Long-Read Sequencing Datasets in Saccharomyces Cerevisiae." *Frontiers in Genetics* 11 (March): 1–10. https://doi.org/10.3389/fgene.2020.00159.

Lupski, James R. 2013. "Genome Mosaicism—One Human, Multiple Genomes." *Science* 341 (July): 358–59.

Luquette, Lovelace J., Craig L. Bohrson, Max A. Sherman, and Peter J. Park. 2019. "Identification of Somatic Mutations in Single Cell DNA-Seq Using a Spatial Model of Allelic Imbalance." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-11857-8.

Magerus-Chatinet, Aude, Bénédicte Neven, Marie-Claude Stolzenberg, Cécile Daussy, Peter D. Arkwright, Nina Lanzarotti, Catherine Schaffner, et al. 2011. "Onset of Autoimmune Lymphoproliferative Syndrome (ALPS) in Humans as a Consequence of Genetic Defect Accumulation." *Journal of Clinical Investigation* 121 (1): 106–12. https://doi.org/10.1172/JCI43752.

Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. "Structural Variant Calling: The Long and the Short of It." *Genome Biology* 20 (1): 1–14. https://doi.org/10.1186/s13059-019-1828-7.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. "The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations." *Nature* 538 (7624): 201–6. https://doi.org/10.1038/nature18964.

Mantere, Tuomo, Simone Kersten, and Alexander Hoischen. 2019. "Long-Read Sequencing Emerging in Medical Genetics." *Frontiers in Genetics* 10 (MAY): 1–14. https://doi.org/10.3389/fgene.2019.00426.

Manthiram, Kalpana, Qing Zhou, Ivona Aksentijevich, and Daniel L. Kastner. 2017. "The Monogenic Autoinflammatory Diseases Define New Pathways in Human Innate Immunity and Inflammation." *Nature Immunology* 18 (8): 832–42. https://doi.org/10.1038/ni.3777.

Martincorena, Iñigo, and Peter J. Campbell. 2015. "Somatic Mutation in Cancer and Normal Cells." *Science* 349 (6255): 1483–89. https://doi.org/10.1126/science.aab4082.

Martincorena, Iñigo, Joanna C. Fowler, Agnieszka Wabik, Andrew R.J. Lawson, Federico Abascal, Michael W.J. Hall, Alex Cagan, et al. 2018. "Somatic Mutant Clones Colonize the Human Esophagus with Age." *Science* 362 (6417): 911–17. https://doi.org/10.1126/science.aau3879.

Martincorena, Iñigo, Amit Roshan, Moritz Gerstung, Peter Ellis, Peter Van Loo, Stuart McLaren, David C. Wedge, et al. 2015. "High Burden and Pervasive Positive Selection of Somatic Mutations in Normal Human Skin." *Science* 348 (6237): 880–86. https://doi.org/10.1126/science.aaa6806.

Martínez-Jiménez, Francisco, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, et al. 2020. "A Compendium of Mutational Cancer Driver Genes." *Nature Reviews Cancer*. https://doi.org/10.1038/s41568-020-0290-x.

Martorana, Davide, Francesco Bonatti, Paola Mozzoni, Augusto Vaglio, and Antonio Percesepe. 2017. "Monogenic Autoinflammatory Diseases with Mendelian Inheritance: Genes, Mutations, and Genotype/Phenotype Correlations." *Frontiers in Immunology* 8 (APR): 1–17. https://doi.org/10.3389/fimmu.2017.00344.

Matsunaga, Takeshi, and Arman Rahman. 1998. "What Brought the Adaptive Immune System to Vertebrates? - The Jaw Hypothesis and the Seahorse." *Immunological Reviews* 166 (2): 177–86. https://doi.org/10.1111/j.1600-065X.1998.tb01262.x.

McDermott, Michael F., Ivona Aksentijevich, Jérôme Galon, Elizabeth M. McDermott, B. William Ogunkolade, Michael Centola, Elizabeth Mansfield, et al. 1999. "Germline Mutations in the Extracellular Domains of the 55 KDa TNF Receptor, TNFR1, Define a Family of Dominantly Inherited Autoinflammatory Syndromes." *Cell* 97 (1): 133–44. https://doi.org/10.1016/S0092-8674(00)80721-7.

McKee, Chloe M., and Rebecca C. Coll. 2020. "NLRP3 Inflammasome Priming: A Riddle Wrapped in a Mystery inside an Enigma." *Journal of Leukocyte Biology* 108 (3): 937–52. https://doi.org/10.1002/JLB.3MR0720-513R.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R.S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1): 1–14. https://doi.org/10.1186/s13059-016-0974-4.

Medzhitov, Ruslan. 2008. "Origin and Physiological Roles of Inflammation." *Nature* 454 (7203): 428–35. https://doi.org/10.1038/nature07201.

Medzhitov, Ruslan, and Charles Jr Janeway. 2000. "The Immune System." Edited by Ian R. Mackay and Fred S. Rosen. *New England Journal of Medicine* 343 (2): 108–17. https://doi.org/10.1056/NEJM200007133430207.

Meienberg, Janine, Rémy Bruggmann, Konrad Oexle, and Gabor Matyas. 2016. "Clinical Sequencing: Is WGS the Better WES?" *Human Genetics* 135 (3): 359–62. https://doi.org/10.1007/s00439-015-1631-9.

Mensa-Vilaró, Anna, María Bravo García-Morato, Oscar de la Calle-Martin, Clara Franco-Jarava, María Teresa Martínez-Saavedra, Luis I. González-Granado, Eva González-Roca, et al. 2018. "Unexpected Relevant Role of Gene Mosaicism in Primary Immunodeficiency Diseases." *Journal of Allergy and Clinical Immunology*, 1–10. https://doi.org/10.1016/j.jaci.2018.09.009.

Mensa-Vilaro, Anna, Weng Tarng Cham, Swee Ping Tang, Sern Chin Lim, Eva González-Roca, Estibaliz Ruiz-Ortiz, Roziana Ariffin, Jordi Yagüe, and Juan I. Aróstegui. 2016. "First Identification of Intrafamilial Recurrence of Blau Syndrome Due to Gonosomal NOD2 Mosaicism." *Arthritis and Rheumatology* 68 (4): 1039–44. https://doi.org/10.1002/art.39519.

Meyer, Matthias, and Martin Kircher. 2010. "Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing." *Cold Spring Harbor Protocols* 5 (6). https://doi.org/10.1101/pdb.prot5448.

Miceli-Richard, Corinne, Suzanne Lesage, Michel Rybojad, Anne Marie Prieur, Sylvie Manouvrier-Hanu, Renate Häfner, Mathias Chamaillard, Habib Zouali, Gilles Thomas, and Jean Pierre Hugot. 2001. "CARD15 Mutations in Blau Syndrome." *Nature Genetics* 29 (1): 19–20. https://doi.org/10.1038/ng720.

Moldavan, A. 1934. "PHOTO-ELECTRIC TECHNIQUE FOR THE COUNTING OF MICROSCOPICAL CELLS." *Science* 80 (2069): 188–89. https://doi.org/10.1126/science.80.2069.188.

Moore, Luiza, Daniel Leongamornlert, Tim H.H. Coorens, Mathijs A. Sanders, Peter Ellis, Stefan C. Dentro, Kevin J. Dawson, et al. 2020. "The Mutational Landscape of Normal Human Endometrial Epithelium." *Nature* 580 (7805): 640–46. https://doi.org/10.1038/s41586-020-2214-z.

Mu, Wenbo, Bing Li, Sitao Wu, Jefferey Chen, Divya Sain, Dong Xu, Mary Helen Black, et al. 2019. "Detection of Structural Variation Using Target Captured Next-Generation Sequencing Data for Genetic Diagnostic Testing." *Genetics in*

Medicine 21 (7): 1603–10. https://doi.org/10.1038/s41436-018-0397-6.

Muckle, Thomas J., and Michael Wells. 1962. "URTICARIA, DEAFNESS, AND AMYLOIDOSIS: A NEW HEREDO-FAMILIAL SYNDROME." *QJM: An International Journal of Medicine* 31 (2): 235–48. https://doi.org/10.1093/oxfordjournals.qjmed.a066967.

Murphy, Kenneth, and Casey Weaver. 2017. *Janeway's Immunobiology. New York: Garland Science.*

Nakamura, K., H. Okamura, M. Wada, K. Nagata, and T. Tamura. 1989. "Endotoxin-Induced Serum Factor That Stimulates Gamma Interferon Production." *Infection and Immunity* 57 (2): 590–95. https://doi.org/10.1128/iai.57.2.590-595.1989.

Nattestad, Maria, and Michael C. Schatz. 2016. "Assemblytics: A Web Analytics Tool for the Detection of Variants from an Assembly." *Bioinformatics* 32 (19): 3021–23. https://doi.org/10.1093/bioinformatics/btw369.

Nicholson, Anna M., Cora Olpe, Alice Hoyle, Ann Sofie Thorsen, Teja Rus, Mathilde Colombé, Roxanne Brunton-Sim, et al. 2018. "Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium." *Cell Stem Cell* 22 (6): 909-918.e8. https://doi.org/10.1016/j.stem.2018.04.020.

Nigrovic, Peter A., Pui Y. Lee, and Hal M. Hoffman. 2020. "Monogenic Autoinflammatory Disorders: Conceptual Overview, Phenotype, and Clinical Approach." *Journal of Allergy and Clinical Immunology* 146 (5): 925–37. https://doi.org/10.1016/j.jaci.2020.08.017.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, et al. 2021. "The Complete Sequence of a Human Genome." *BioRxiv*, January, 2021.05.26.445798. https://doi.org/10.1101/2021.05.26.445798.

Odagiri, Y., H. Uchida, M. Hosokawa, K. Takemoto, A. A. Morley, and T. Takeda. 1998. "Accelerated Accumulation of Somatic Mutations in the Senescence-Accelerated Mouse." *Nature Genetics* 19 (2): 116–17. https://doi.org/10.1038/468.

Osorio, Fernando G., Axel Rosendahl Huber, Rurika Oka, Mark Verheul, Sachin H. Patel, Karlijn Hasaart, Lisanne de la Fonteijne, Ignacio Varela, Fernando D. Camargo, and Ruben van Boxtel. 2018. "Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis." *Cell Reports* 25 (9): 2308-2316.e4. https://doi.org/10.1016/j.celrep.2018.11.014.

Pamcer, Zeev, Chris T. Amemiya, Götz R.A. Ehrhardt, Jill Coitlin, G. Larry Gartland, and Max D. Cooper. 2004. "Somatic Diversification of Variable Lymphocyte Receptors in the Agnathan Sea Lamprey." *Nature* 430 (6996): 174–80. https://doi.org/10.1038/nature02740.

Pareek, Chandra Shekhar, Rafal Smoczynski, and Andrzej Tretyn. 2011. "Sequencing Technologies and Genome Sequencing." *Journal of Applied Genetics* 52 (4): 413–35. https://doi.org/10.1007/s13353-011-0057-x.

Pich, Oriol, Iker Reyes-Salazar, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. 2020. "Discovering the Drivers of Clonal Hematopoiesis." *BioRxiv*, 2020.10.22.350140. https://doi.org/10.1101/2020.10.22.350140.

Picot, Julien, Coralie L. Guerin, Caroline Le Van Kim, and Chantal M. Boulanger. 2012. "Flow Cytometry: Retrospective, Fundamentals and Recent Instrumentation." *Cytotechnology* 64 (2): 109–30. https://doi.org/10.1007/s10616-011-9415-0.

Pollard, Martin O., Deepti Gurdasani, Alexander J. Mentzer, Tarryn Porter, and Manjinder S. Sandhu. 2018. "Long Reads: Their Purpose and Place." *Human*

Molecular Genetics 27 (R2): R234–41. https://doi.org/10.1093/hmg/ddy177.

Rang, Franka J., Wigard P. Kloosterman, and Jeroen de Ridder. 2018. "From Squiggle to Basepair: Computational Approaches for Improving Nanopore Sequencing Read Accuracy." *Genome Biology* 19 (1): 1–11. https://doi.org/10.1186/s13059-018-1462-9.

Reddy, Sreelatha, Shuang Jia, Rhonda Geoffrey, Rachel Lorier, Mariko Suchi, Ulrich Broeckel, Martin J. Hessner, and James Verbsky. 2009. "An Autoinflammatory Disease Due to Homozygous Deletion of the IL1RN Locus." *New England Journal of Medicine* 360 (23): 2438–44. https://doi.org/10.1056/NEJMoa0809568.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. "Global Variation in Copy Number in the Human Genome." *Nature* 444 (7118): 444–54. https://doi.org/10.1038/nature05329.

Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. "CADD: Predicting the Deleteriousness of Variants throughout the Human Genome." *Nucleic Acids Research* 47 (D1): D886–94. https://doi.org/10.1093/nar/gky1016.

Rhoads, Anthony, and Kin Fai Au. 2015. "PacBio Sequencing and Its Applications." *Genomics, Proteomics and Bioinformatics* 13 (5): 278–89. https://doi.org/10.1016/j.gpb.2015.08.002.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. "Integrative Genomics Viewer." *Nature Biotechnology* 29 (1): 24–26. https://doi.org/10.1038/nbt.1754.

Rowczenio, Dorota M., Hadija Trojer, Ebun Omoyinmi, Juan I. Aróstegui, Grigor Arakelov, Anna Mensa-Vilaro, Anna Baginska, et al. 2016. "Brief Report: Association of Tumor Necrosis Factor Receptor–Associated Periodic Syndrome With Gonosomal Mosaicism of a Novel 24-Nucleotide TNFRSF1A Deletion." *Arthritis and Rheumatology* 68 (8): 2044–49. https://doi.org/10.1002/art.39683.

Sandmann, Sarah, Aniek O. De Graaf, Mohsen Karimi, Bert A. Van Der Reijden, Eva Hellström-Lindberg, Joop H. Jansen, and Martin Dugas. 2017. "Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data." *Scientific Reports* 7: 1–12. https://doi.org/10.1038/srep43169.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences* 74 (12): 5463–67. https://doi.org/10.1073/pnas.74.12.5463.

Santer, R, R Schneppenheim, A Dombrowski, H Gotze, B Steinmann, and J Schaub. 1997. "A Candidate Gene for Familial Mediterranean Fever." *Nature Genetics* 15: 57–61.

Seckinger, P, J W Lowenthal, K Williamson, J M Dayer, and H R MacDonald. 1987. "A Urine Inhibitor of Interleukin 1 Activity That Blocks Ligand Binding." *Journal of Immunology* 139 (5): 1546–49. http://www.ncbi.nlm.nih.gov/pubmed/2957429.

Sedlazeck, Fritz J., Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. 2018. "Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing." *Nature Methods* 15 (6): 461–68. https://doi.org/10.1038/s41592-018-0001-7.

Shinar, Yael, Tali Tohami, Avi Livneh, Ginette Schiby, Abraham Hirshberg, Meital Nagar, Itamar Goldstein, et al. 2015. "Acquired Familial Mediterranean Fever

Associated with a Somatic MEFV Mutation in a Patient with JAK2 Associated Post-Polycythemia Myelofibrosis." *Orphanet Journal of Rare Diseases* 10 (1): 1–6. https://doi.org/10.1186/s13023-015-0298-6.

Shiraishi, Yuichi, Yusuke Sato, Kenichi Chiba, Yusuke Okuno, Yasunobu Nagata, Kenichi Yoshida, Norio Shiba, et al. 2013. "An Empirical Bayesian Framework for Somatic Mutation Detection from Cancer Genome Sequencing Data." *Nucleic Acids Research* 41 (7). https://doi.org/10.1093/nar/gkt126.

Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. "Detecting DNA Cytosine Methylation Using Nanopore Sequencing." *Nature Methods* 14 (4): 407–10. https://doi.org/10.1038/nmeth.4184.

Sims, John E., and Dirk E. Smith. 2010. "The IL-1 Family: Regulators of Immunity." *Nature Reviews Immunology* 10 (2): 89–102. https://doi.org/10.1038/nri2691.

Smith, Kyle S., Vinod K. Yadav, Shanshan Pei, Daniel A. Pollyea, Craig T. Jordan, and Subhajyoti De. 2016. "SomVarIUS: Somatic Variant Identification from Unpaired Tissue Samples." *Bioinformatics* 32 (6): 808–13. https://doi.org/10.1093/bioinformatics/btv685.

Song, Kelly. 2018. "4 Risks Consumers Need to Know about DNA Testing Kit Results and Buying Life Insurance." *CNBC*, 2018. https://www.cnbc.com/2018/08/04/4--risks-consumer-face-with-dna-testing-and-buying-life-insurance.html.

Spinney, Laura. 2020. "Your DNA Is a Valuable Asset, so Why Give It to Ancestry Websites for Free?" *The Guardian*, 2020. https://www.theguardian.com/commentisfree/2020/feb/16/dna-hugely-valuable-health-tech-privacy.

Starokadomskyy, Petro, Terry Gemelli, Jonathan J. Rios, Chao Xing, Richard C Wang, Haiying Li, Vladislav Pokatayev, et al. 2016. "DNA Polymerase-α Regulates the Activation of Type I Interferons through Cytosolic RNA:DNA Synthesis." *Nature Immunology* 17 (5): 495–504. https://doi.org/10.1038/ni.3409.

Steiner, Annemarie, Cassandra R. Harapas, Seth L. Masters, and Sophia Davidson. 2018. "An Update on Autoinflammatory Diseases: Relopathies." *Current Rheumatology Reports* 20 (7): 39. https://doi.org/10.1007/s11926-018-0749-x.

Sun, Yu, Jiale Xiang, Yidong Liu, Sen Chen, Jintao Yu, Jiguang Peng, Zijing Liu, et al. 2019. "Increased Diagnostic Yield by Reanalysis of Data from a Hearing Loss Gene Panel." *BMC Medical Genomics* 12 (1): 1–8. https://doi.org/10.1186/s12920-019-0531-6.

Sutterwala, Fayyaz S., Stefanie Haasken, and Suzanne L. Cassel. 2014. "Mechanism of NLRP3 Inflammasome Activation." *Annals of the New York Academy of Sciences* 1319 (1): 82–95. https://doi.org/10.1111/nyas.12458.

Tang, Jessica, Eleanor Fewings, Darwin Chang, Hanlin Zeng, Shanshan Liu, Aparna Jorapur, Rachel Belote, et al. 2020. "The Genomic Landscapes of Individual Melanocytes from Human Skin." *Nature*, no. April. https://doi.org/10.1101/2020.03.01.971820.

Tangye, Stuart G, Waleed Al-Herz, Aziz Bousfiha, Talal Chatila, Charlotte Cunningham-Rundles, Amos Etzioni, Jose Luis Franco, et al. 2020. "Human Inborn Errors of Immunity: 2019 Update on the Classification from the International Union of Immunological Societies Expert Committee." *Journal of Clinical Immunology* 40 (1): 24–64. https://doi.org/10.1007/s10875-019-00737-x.

Tattini, Lorenzo, Romina D'Aurizio, and Alberto Magi. 2015. "Detection of Genomic

Structural Variants from Next-Generation Sequencing Data." *Frontiers in Bioengineering and Biotechnology* 3 (JUN): 1–8. https://doi.org/10.3389/fbioe.2015.00092.

Teer, Jamie K., Yonghong Zhang, Lu Chen, Eric A. Welsh, W. Douglas Cress, Steven A. Eschrich, and Anders E. Berglund. 2017. "Evaluating Somatic Tumor Mutation Detection without Matched Normal Samples." *Human Genomics* 11 (1): 1–13. https://doi.org/10.1186/s40246-017-0118-2.

The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Touitou, Isabelle, Jean Marc Rey, Christiane Dross, Madeleine Dupont, Olivier Brun, Marc Ciano, Jacques Demaille, et al. 1996. "Localization of the Familial Mediterranean Fever Gene (FMF) to a 250-Kb Interval in Non-Ashkenazi Jewish Founder Haplotypes." *American Journal of Human Genetics* 59 (3): 603–12.

Uhlén, Mathias, Linn Fagerberg, Bjö M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Tissue-Based Map of the Human Proteome." *Science* 347 (6220). https://doi.org/10.1126/science.1260419.

Valles-Ibáñez, Guillem de, Ana Esteve-Solé, Mònica Piquer, E. Azucena González-Navarro, Jessica Hernandez-Rodriguez, Hafid Laayouni, Eva González-Roca, et al. 2018. "Evaluating the Genetics of Common Variable Immunodeficiency: Monogenetic Model and Beyond." *Frontiers in Immunology* 9 (MAY): 1–15. https://doi.org/10.3389/fimmu.2018.00636.

Vaser, Robert, Swarnaseetha Adusumalli, Sim Ngak Leng, Mile Sikic, and Pauline C. Ng. 2016. "SIFT Missense Predictions for Genomes." *Nature Protocols* 11 (1): 1–9. https://doi.org/10.1038/nprot.2015.123.

Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51. https://doi.org/10.1126/science.1058040.

Wang, Yifan, Taejeong Bae, Jeremy Thorpe, Maxwell A. Sherman, Attila G. Jones, Sean Cho, Kenneth Daily, et al. 2021. "Comprehensive Identification of Somatic Nucleotide Variants in Human Brain Tissue." *Genome Biology* 22 (1): 1–32. https://doi.org/10.1186/s13059-021-02285-3.

Watkin, Levi B., Birthe Jessen, Wojciech Wiszniewski, Timothy J. Vece, Max Jan, Youbao Sha, Maike Thamsen, et al. 2015. "COPA Mutations Impair ER-Golgi Transport and Cause Hereditary Autoimmune-Mediated Lung Disease and Arthritis." *Nature Genetics* 47 (6): 654–60. https://doi.org/10.1038/ng.3279.

Watson, Caroline J, A L Papula, Gladys Y P Poon, Wing H Wong, and Andrew L Young. 2020. "The Evolutionary Dynamics and Fitness Landscape of Clonal Hematopoiesis." *Science* 1454 (March): 1449–54.

Welch, John S., Timothy J. Ley, Daniel C. Link, Christopher A. Miller, David E. Larson, Daniel C. Koboldt, Lukas D. Wartman, et al. 2012. *The Origin and Evolution of Mutations in Acute Myeloid Leukemia*. *Cell.* Vol. 150. https://doi.org/10.1016/j.cell.2012.06.023.

Wells, Alex, David Heckerman, Ali Torkamani, Li Yin, Jonathan Sebat, Bing Ren, Amalio Telenti, and Julia di Iulio. 2019. "Ranking of Non-Coding Pathogenic Variants and Putative Essential Regions of the Human Genome." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-13212-3.

Werling, Donna M., Harrison Brand, Joon-Yong An, Matthew R. Stone, Lingxue Zhu, Joseph T. Glessner, Ryan L. Collins, et al. 2018. "An Analytical Framework for

Whole-Genome Sequence Association Studies and Its Implications for Autism Spectrum Disorder." *Nature Genetics* 50 (5): 727–36. https://doi.org/10.1038/s41588-018-0107-y.

Whitford, Whitney, Klaus Lehnert, Russell G. Snell, and Jessie C. Jacobsen. 2019. "Evaluation of the Performance of Copy Number Variant Prediction Tools for the Detection of Deletions from Whole Genome Sequencing Data." *Journal of Biomedical Informatics* 94 (December 2018): 103174. https://doi.org/10.1016/j.jbi.2019.103174.

Wilm, Andreas, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. 2012. "LoFreq: A Sequence-Quality Aware, Ultra-Sensitive Variant Caller for Uncovering Cell-Population Heterogeneity from High-Throughput Sequencing Datasets." *Nucleic Acids Research* 40 (22): 11189–201. https://doi.org/10.1093/nar/gks918.

Won, Dongju, Se Hee Kim, Borahm Kim, Seung Tae Lee, Hoon Chul Kang, and Jong Rak Choi. 2020. "Reanalysis of Genomic Sequencing Results in a Clinical Laboratory: Advantages and Limitations." *Frontiers in Neurology* 11 (June): 1–6. https://doi.org/10.3389/fneur.2020.00612.

Wright, Caroline F., Jeremy F. McRae, Stephen Clayton, Giuseppe Gallone, Stuart Aitken, Tomas W. FitzGerald, Philip Jones, et al. 2018. "Making New Genetic Diagnoses with Old Data: Iterative Reanalysis and Reporting from Genome-Wide Data in 1,133 Families with Developmental Disorders." *Genetics in Medicine* 20 (10): 1216–23. https://doi.org/10.1038/gim.2017.246.

Xie, Mingchao, Charles Lu, Jiayin Wang, Michael D. McLellan, Kimberly J. Johnson, Michael C. Wendl, Joshua F. McMichael, et al. 2014. "Age-Related Mutations Associated with Clonal Hematopoietic Expansion and Malignancies." *Nature Medicine* 20 (12): 1472–78. https://doi.org/10.1038/nm.3733.

Xing, Dong, Longzhi Tan, Chi-Han Chang, Heng Li, and X. Sunney Xie. 2021. "Accurate SNV Detection in Single Cells by Transposon-Based Whole-Genome Amplification of Complementary Strands." *Proceedings of the National Academy of Sciences* 118 (8): e2013106118. https://doi.org/10.1073/pnas.2013106118.

Xu, Chang, Xiujing Gu, Raghavendra Padmanabhan, Zhong Wu, Quan Peng, John DiCarlo, and Yexun Wang. 2019. "SmCounter2: An Accurate Low-Frequency Variant Caller for Targeted Sequencing Data with Unique Molecular Identifiers." Edited by Inanc Birol. *Bioinformatics* 35 (8): 1299–1309. https://doi.org/10.1093/bioinformatics/bty790.

Xu, Huilei, John DiCarlo, Ravi V. Satya, Quan Peng, and Yexun Wang. 2014. "Comparison of Somatic Mutation Calling Methods in Amplicon and Whole Exome Sequence Data." *BMC Genomics* 15 (1): 1–10. https://doi.org/10.1186/1471-2164-15-244.

Yang, Xiaolu., Howard Y. Chang, and David Baltimore. 1998. "Autoproteolytic Activation of Pro-Caspases by Oligomerization." *Molecular Cell* 1 (2): 319–25. https://doi.org/10.1016/S1097-2765(00)80032-5.

Yang, Yaping, Donna M. Muzny, Jeffrey G. Reid, Matthew N Bainbridge, Alecia Willis, Patricia A Ward, Alicia Braxton, et al. 2013. "Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders." *New England Journal of Medicine* 369 (16): 1502–11. https://doi.org/10.1056/NEJMoa1306555.

Yasuda, Koubun, Kenji Nakanishi, and Hiroko Tsutsui. 2019. "Interleukin-18 in Health and Disease." *International Journal of Molecular Sciences* 20 (3).

https://doi.org/10.3390/ijms20030649.

Yazici, Hasan, Emire Seyahi, Gulen Hatemi, and Yusuf Yazici. 2018. "Behçet Syndrome: A Contemporary View." *Nature Reviews Rheumatology* 14 (2): 107–19. https://doi.org/10.1038/nrrheum.2017.208.

Yim, Seon Hee, Seung Hyun Jung, Boram Chung, and Yeun Jun Chung. 2015. "Clinical Implications of Copy Number Variations in Autoimmune Disorders." *Korean Journal of Internal Medicine* 30 (3): 294–304. https://doi.org/10.3904/kjim.2015.30.3.294.

Yokoyama, Akira, Nobuyuki Kakiuchi, Tetsuichi Yoshizato, Yasuhito Nannya, Hiromichi Suzuki, Yasuhide Takeuchi, Yusuke Shiozawa, et al. 2019. "Age-Related Remodelling of Oesophageal Epithelia by Mutated Cancer Drivers." *Nature* 565 (7739): 312–17. https://doi.org/10.1038/s41586-018-0811-x.

Yoshida, Kenichi, Kate H.C. Gowers, Henry Lee-Six, Deepak P. Chandrasekharan, Tim Coorens, Elizabeth F. Maughan, Kathryn Beal, et al. 2020. "Tobacco Smoking and Somatic Mutations in Human Bronchial Epithelium." *Nature* 578 (7794): 266–72. https://doi.org/10.1038/s41586-020-1961-1.

Zhang, Feng., and James R. Lupski. 2015. "Non-Coding Genetic Variants in Human Disease." *Human Molecular Genetics* 24 (R1): R102–10. https://doi.org/10.1093/hmg/ddv259.

Zhang, Le, Wanyu Bai, Na Yuan, and Zhenglin Du. 2019. "Comprehensively Benchmarking Applications for Detecting Copy Number Variation." Edited by Ilya Ioshikhes. *PLOS Computational Biology* 15 (5): e1007069. https://doi.org/10.1371/journal.pcbi.1007069.

Zhao, Min, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. 2013. "Computational Tools for Copy Number Variation (CNV) Detection Using next-Generation Sequencing Data: Features and Perspectives." *BMC Bioinformatics* 14 (S11): S1. https://doi.org/10.1186/1471-2105-14-S11-S1.

Zheng, Jianchao, Hongyun Zhang, Santasree Banerjee, Yun Li, Junyu Zhou, Qian Yang, Xuemei Tan, et al. 2019. "A Comprehensive Assessment of Next-Generation Sequencing Variants Validation Using a Secondary Technology." *Molecular Genetics & Genomic Medicine* 7 (7): 1–7. https://doi.org/10.1002/mgg3.748.

Zhou, Anbo, Timothy Lin, and Jinchuan Xing. 2019. "Evaluating Nanopore Sequencing Data Processing Pipelines for Structural Variation Identification." *Genome Biology* 20 (1): 1–13. https://doi.org/10.1186/s13059-019-1858-1.

Zink, Florian, Simon N. Stacey, Gudmundur L. Norddahl, Michael L. Frigge, Olafur T. Magnusson, Ingileif Jonsdottir, Thorgeir E. Thorgeirsson, et al. 2017. "Clonal Hematopoiesis, with and without Candidate Driver Mutations, Is Common in the Elderly." *Blood* 130 (6): 742–52. https://doi.org/10.1182/blood-2017-02-769869.