Population genetics of Copy Number Variants: the case of the Romani population

Marco Antinucci

TESI DOCTORAL UPF / 2023

Thesis supervisor

Dr. Francesc Calafell

DEPARTMENT OF MEDICINE AND LIFE SCIENCES

Universitat Pompeu Fabra Barcelona

"*Looming, omnipresent, this task ahead this task at hand.*
Ominous and daunting crippling undertaking, I'm frozen.
Where to begin eludes me without you to remind me […]
Take the step, take the swing, take the bite just take the bite
Take the step, take the swing, take the bite, just go all in.
Where to begin eludes me, without you to remind me: just begin."
A Perfect Circle – Eat the Elephant

Acknowledgements

Deseo agradecer Francesc que desde aquel verano del 2017 empezó a interesarse a mi solicitud para empezar un doctorado y en los primeros mails dejó claro que no hacía falta ser tan formal y seguir dándole “del Lei”. Gracias por haberme dado esta ocasión de seguir trabajando en la investigación con este proyecto y creer en mí, que venía de un entorno más pequeño y con experiencia previa en proyectos menos ambiciosos. No ha sido fácil superar todos los problemas y llegar al final, pero lo hemos conseguido, gracias por ayudarme en los momentos más difíciles y por tu soporte.

Carla, Simone, André, Barbara, Neus, Rocío, Ana y Marcel. Chicos, que decir… Habéis sido una multitud de cosas para mí y entre nosotros: compañeros de trabajo hablando de los proyectos y ayudándonos a solucionar problemas, profesionales psicólogos en los grupos de terapia colectiva e individual, amigos para quedar y olvidarse de todo en un día malo. He tenido la suerte de encontrar un entorno familiar que ha sabido escuchar, aconsejar y ha compartido conmigo sus momentos personales, estoy contento de haberos conocido y que nos hemos acompañado en un tramo tan largo (y corto a la vez) de vida.

Respecto a las personas que entraron después de mí, Julen, Jose, Laura, Simón, Jorge y Marcos, he compartido menos tiempo con vosotros respecto a los “antiguos” pero gracias por las charlas y las risas (¡Paris fue genial!), sois la nueva generación del IBE, que se queda en buenas manos. Harvi, thanks for your kindness and for being always positive about what I had to do next, for all the support (technical and not) and for sharing our experiences during these years. Judit, escribo pocas palabras así no te quejas de que se hace demasiada larga esa dedica; además del soporte que das a todos, gracias por estar siempre lista para ir más allá de los temas del trabajo y tener unas palabras y opiniones sobre las cosas que nos pasan.

Chiara, finalmente sembra che questo dottorato sia arrivato alla fine. Non sono stati anni facili, fatti di distanze, programmazione degli incontri successivi e tante ore negli aeroporti. Grazie per aver creduto in quello che facevo, per avermi incoraggiato nei tanti periodi difficili e ricordato che ho un valore che non viene solo misurato dall'esito del lavoro che faccio. Sei la mia ancora durante le tempeste e il porto sicuro verso il quale tornare. Ho imparato a dare ancora più valore ai momenti passati insieme e insieme siamo riusciti a dare minor peso al tempo trascorso dall'ultimo incontro, concentrandoci di più su quello rimanente al successivo. Grazie per l'amore che mi dimostri dopo tutto questo tempo, sono felice e fortunato ad avere te.

Un pensiero agli amici Pisani, e non solo, che mi hanno sostenuto anche solo facendo quattro chiacchiere, grazie a Riccardo (import pandas as pd), Francesco, Ania, Lorenzo, Lisa, Marco, Licia, Giorgio, Davide e Anna. Grazie a Carlo per gli innumerevoli quanto grandi aiuti informatici (indispensabili in alcuni frangenti) e la buona dose di conversazioni su temi più disparati, buona fortuna col proseguimento del tuo dottorato.

Grazie alla mia famiglia, a mio padre che gestisce le "cose di casa" ed è il tassista di famiglia ufficiale quando arrivo o parto, a Chiara, Stefano e Irene per essere sempre presenti quando torno a casa e a compensare la distanza da lontano con (video)chiamate, messaggi vocali e "foffi". Un abbraccio anche a Bruna e Mauro che hanno sempre un pensiero per me.

Hai visto fin dove sono arrivato mamma? Probabilmente saresti stata contenta del fatto che abbia seguito le tue orme e dei piccoli risultati raggiunti fin qui, questa tesi è anche per te.

Mirko, non riesco a dire molto senza riaprire le ferite, magari per la discussione di questa tesi sarei riuscito a farti salire su un aereo per Barcellona, grazie per tutti i momenti divertenti e assurdi passati insieme, hanno contribuito a rendermi la persona che sono.

Abstract

Much information has been gathered for the Romani population so far, spanning historical, linguistics and genetics research. While whole genome analyses for this population started only recently, most studies using genetic data so far only relied on single nucleotide polymorphisms (SNPs). In this work we address, for the first time, whole genome Copy Number Variant (CNV) markers in the Romani. Using deletions, we reconstruct the relationships of Romani with related populations (Eurasian and South Asian) and highlight differentiation among them. Moreover, increased presence of deletions at Loss of Function (LoF)-intolerant genes in Romani points to a relaxation of natural selection towards putative slightly deleterious variants in the population. Finally, Romani show overrepresentation of such deletions in genes related to nervous system features and, moreover, their linkage disequilibrium with SNPs in previously reported genes of biomedical importance in Romani could suggest a contribution of CNVs and SNPs to phenotypically related outcomes.

Riassunto

Fino ad oggi, molte informazioni sono state raccolte sul popolo Romanì utilizzando ricerche storiche, linguistiche e genetiche. Sebbene l'analisi di genomi completi in questo popolo sia iniziata recentemente, la maggior parte degli studi genetici si è concentrata sull'utilizzo dei polimorfismi a singolo nucleotide (SNPs). In questo studio analizziamo, per la prima volta, le varianti del numero di copie (CNVs) in sequenze genomiche complete nei Romanì. Utilizzando le delezioni, ricostruiamo le relazioni dei Romanì con popolazioni a loro affini (Euroasiatiche, Sud asiatiche), mostrando anche le differenziazioni tra le suddette. Inoltre, la maggior presenza di delezioni in geni intolleranti alla perdita di funzione nei Romanì indica un rilassamento della selezione naturale verso mutazioni solo lievemente dannose. In ultimo, i Romanì presentano maggiormente questo tipo di delezioni in geni implicati in funzioni relative al sistema nervoso e, inoltre, il loro linkage disquilibirum con SNPs risiedenti in geni di interesse biomedico, già riportati in letteratura nei Romanì, potrebbe indicare un'azione sinergica di CNVs e SNPs verso esiti fenotipicamente correlati.

Preface

Studying human population genetics, thus focussing on the variation at the genetic level among humans from worldwide locations and the processes occurred to produced it, provided many insights for the evolutionary history of *Homo sapiens* and its ancestors. Indeed, during the history towards modern humans as well as after their colonization of the Earth, individuals within populations faced different conditions and stressors from the wide set of environments they inhabited and favourable genetic features in each circumstance were more likely to persist and pass to the next generations. Random changes of allele frequencies play a role in defining the destiny of variants within populations and different populations can experience more or less drastic variations of these distribution changes. Human populations, however, are not closed entities reacting strictly to the environment they experience, instead they move, fragment and mix with each other. All these processes help reshuffling bits of "variability sets" of each contributor, which can have functional consequence (and be more or less prone to pass on subsequent generations) or be neutral but, that if traced back, can reveal the past histories of human groups.

Historically, human population genetics has been addressed predominantly by leveraging single nucleotide polymorphisms (SNPs) as the marker of choice, this allowed a wide evolution and refinement of both detection and inference techniques. Nonetheless, although only recently, the contribution of copy number variants (CNVs) to human differentiation, evolution and health has been largely recognised, sparking the interest in what can be inferred also including this type of variants. CNVs in fact, encompassing larger stretches of nucleotides than SNPs and thus are responsible for a higher amount of nucleotide variation. As a result, their impact could be stronger than point mutations, playing a role in the biology and evolution of humans.

In this PhD thesis the interesting features of CNVs are analysed for the first time within the context of a historically isolated and underrepresented population, the Romani. The study poses its foundations on a double purpose: one is assessing the information that can be detected, using CNVs, in a population whose peculiar

demographic history has been addressed using classical genetic markers; the second is exploiting this new source of information in such an interesting population scenario, to investigate new putative information previously undescribed. The first instance indeed can reveal the potential of CNVs to cover (or not) known details of Romani and allowing an evaluation of the informative power of these markers within an underrepresented group. The second aspect focusses more on the known features of CNVs to explore further the genetic information in Romani that might have escaped research with classical markers, making them suitable and interesting for advancing the knowledge within this context.

TABLE OF CONTENTS

List of figures

List of tables

1. INTRODUCTION

1.1 Structural variants

Most of the studies on human genetics have been conducted typically using single nucleotide variants or polymorphisms (SNVs or SNPs), namely mutations occurring at a single base pair (bp), which differ for their frequency. Indeed, a SNP is an SNV present at least in 1% of the population. Additionally, other types of mutations exist, such as short (<50 bp) insertion or deletions, which indicate gains or losses of genetic material, known as INDELs and structural variants (SVs) which comprehend a generally larger class of variants. While SNPs and INDELs have largely been more studied, SVs have usually lagged behind in research, a research that once had started to address these variants highlighted different interesting features of this class of markers. In this chapter we provide an overview of SVs and information about their formation at a molecular level, their role in evolutionary processes and health, and how research has increasingly refined strategies to detect them.

1.1.1 Overview

The term structural variants refers to differences in the genome architecture of an organism with respect to a reference, usually involving more than 50 bp and up to several million bp (Mb). These variable chunks of DNA can either be newly inserted, removed, duplicated, inverted or relocated elsewhere, giving rise to what the scientific community refers to as insertions, deletions, duplications, inversions and translocations, respectively. Being larger than SNPs and INDELs, SVs contribute to most of the variability in the human genome (Levy et al. 2007; Wheeler et al. 2008; Conrad et al. 2010; Weischenfeldt et al. 2013; Sudmant, Rausch, et al. 2015). Within SVs we can identify balanced and unbalanced mutations. Balanced mutations do not alter the overall content of a genome but rather reorganize portions of DNA spatially, like inversions and balanced translocations. Inversions changes the orientation of a portion of DNA, as

implied by the name, by inverting its sequence. First described and in *Drosophila* species throughout the 20th century (Sturtevant 1921; Dobzhansky 1970; Krimbas and Powell 1992), inversions have been shown to be taxonomically widespread, to reduce recombination around and within them, play a role in reproductive behaviours/isolation and foster adaptive changes within taxa (Kirkpatrick and Barton 2006; Lowry and Willis 2010; Jones et al. 2012; Farré et al. 2013; Küpper et al. 2016; Barth et al. 2017). One of the most common SV types in humans are balanced translocations, namely relocations of genome portions that can happen at a macroscopic scale, such as the fusion of acrocentric chromosomes at centromeres – also known as Robertsonian translocations - or between two non-homologous chromosomes exchanging genomic material with each other (Mack and Swisshelm 2013; Weckselblatt et al. 2015; Morin et al. 2017). Being balanced, such events usually do not usually alter phenotypes, even though humans with identified reciprocal translocations may suffer infertility and pregnancy failures (Morin et al. 2017). Conversely, unbalanced mutations do alter the genome dosage with two opposite processes, gains and losses of genetic material. The former is represented by insertions and duplications, the latter by deletions plus a hybrid source of disruption, namely unbalanced translocation. While insertions are novel nucleotide incorporations within a genetic sequence, duplications involve the copy of a genome portion that is integrated at least once in the genome[1], even though they can also configure in multi-copy duplications. Deletions, on the other hand, work as the opposed mechanism, removing sequence fragments from the genome. Translocations can operate also to produce unbalanced mutations, such as in macroscopic monosomic or trisomic events, accounting for ~1% of known cases of intellectual disability, developmental delay and birth defects (Ravnan 2006; Ballif et al. 2007;

[1] If the copied segments are incorporated adjacent to the original one, we refer to tandem duplications, as opposed to interspersed duplications

Shao et al. 2008; Weckselblatt and Rudd 2015a). Together, variants under the unbalanced category aside from insertions, form their own subclass of SVs, namely Copy Number Variations, or Copy Number Variants (CNVs), as they underlie differences in the amount of genome portions either gained or lost, among individuals (Figure 1).

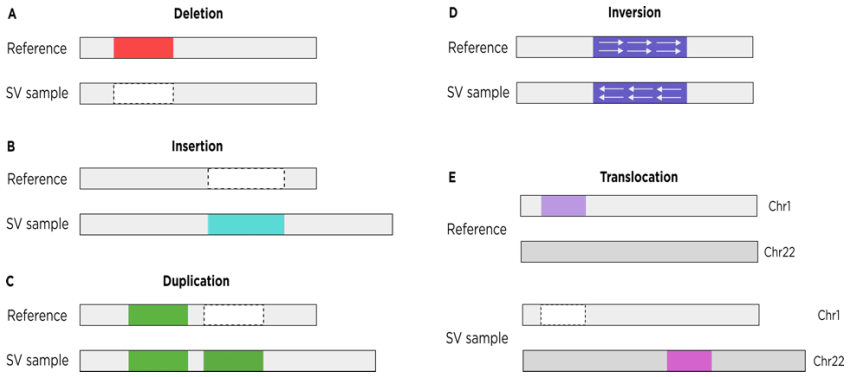


Figure 1. Illustration of structural variant types. Unbalanced SVs, changing overall genome dosage are depicted on the right of the figure (**A**, **B**, **C**) while balanced SVs, leaving the net dosage unaltered are arranged on the right (**D**, **E**). **A**. A deleted segments is missing in the sample with this SV. **B.** A sequence is inserted with respect to a reference. **C**. One portion of the genome is duplicated. **D**. The sequence is inserted back with an inverted orientation with respect to the original position. **E**. A translocation displaces a sequence from a chromosome to another.

Indeed, macroscopic variations in genome contents upon inspection of karyotypes, throughout the second part of the 20th century, revealed structural variations in humans - most often leading to disorders as the size detected ranged from Mb to whole chromosomes in size (Nowell and Hungerford 1960; Craig-Holmes and Shaw 1971; Zankl and Zang 1971; Rowley 1973). The turn of the century brought new methodologies to address structural variation and studies started accumulating knowledge on smaller SVs in healthy humans, not detectable using previous macroscopic approaches (Iafrate et al. 2004a; Sebat et al. 2004; Feuk et al. 2006). An increasing amount of CNVs in individuals with benign and disease phenotypes was further found, and a common substrate of variation among humans was uncovered, emphasizing the general idea that SVs have a role

on phenotypic traits, gene expression and the evolutionary implications of their diversification for the selective processes acting upon them (Dhami et al. 2005; Tuzun et al. 2005; Tyson et al. 2005; McCarroll et al. 2006; Stranger et al. 2007; Hurles et al. 2008; Perry et al. 2008).

1.1.2 Mechanisms of formation

As SNVs are the result of point mutations that occur due to replication errors or imperfect repair mechanisms after damage, also SVs form under specific circumstances and mechanisms that play a significant role in both their frequency of formation and type of constitutional alteration. As the former type of genetic alterations, SVs can form both in the somatic line or in the germline, in the first case even though somatic SVs do not strictly imply the presence of a tumour, they are usually found in the aberrant genomic architectures of cancer genomes. Germline SVs, as for other types of genetic mutations, produce structural changes that do not jeopardize the stability of the cell they reside, they may pass to the subsequent generation, and give rise to polymorphisms among individuals and functional implications. The main background mechanisms responsible for SVs formation, as stated above for single nucleotide events, involve repair processes, recombination events and DNA replication (Carvalho and Lupski 2016) and an additional intriguing detail for the occurrence of structural changes is that genomic architecture itself may result in DNA regions more prone to such types of alterations (Shaw 2004). More specifically, most common mechanisms leading to the formation of SVs are initiated after a double strand break (DSB) or replication fork stalling/disruption, both leading to recombination or incorrect repair. These strand breaks (Figure 2), which can appear along a common DNA segment (double-ended) or in one of the two double strand segments at a replication fork (single-ended) (Piazza and Heyer 2019) can be caused by environmental stressors in the cell such as different ionizing radiations,

reactive oxygen species and mechanical forces (Gu et al. 2008; Lieber 2010).

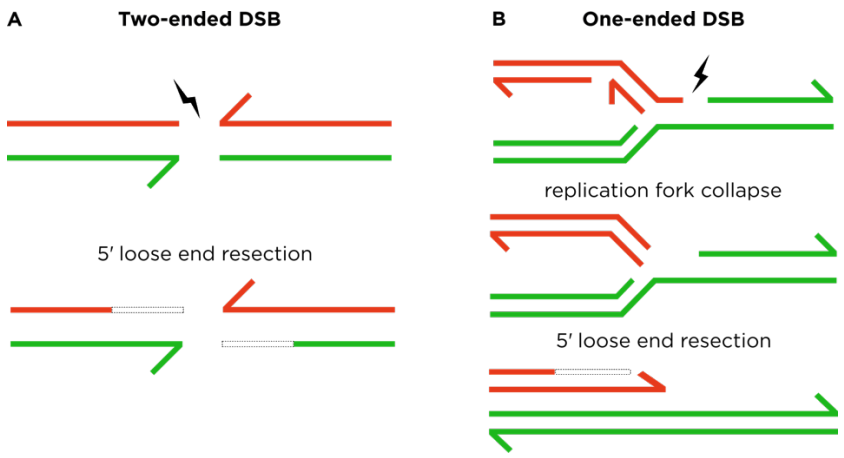


Figure 2. **Double strand break**. The figure shows double strand break (DSB), lines of different colours represent DNA chains, the bolt icon represents the damaged point, the small spines at the end of the strands represent the 3' ends. **A**. Two-ended DSB. A damage snips the DNA strand, after the damage, the 5' loose portion of the chain is resected, these ends can subsequently interact with other strands. **B**. One-ended DSB. A damage nicks the lagging strand, causing the collapse of the replication fork and interruption of replication on the leading strand, followed by elimination of nucleotides on the 5'end. Repair mechanisms subsequently acts in different ways to repair the damage.

Replication fork faults instead can be caused by interaction with RNA molecules, proteins and DNA lacerations (Hastings, Ira, et al. 2009; Hattori and Fukami 2020) but also by uncommon DNA structures (Figure 3) in the presence of repetitive sequences. Indeed, inverted and mirror repeats can create harpins and triple-strand structures (H-DNA) respectively, while direct G-rich tandem repeats result in G-quadruplex all disturbing DNA replication machinery (Lee et al. 2007a; Mirkin and Mirkin 2007).

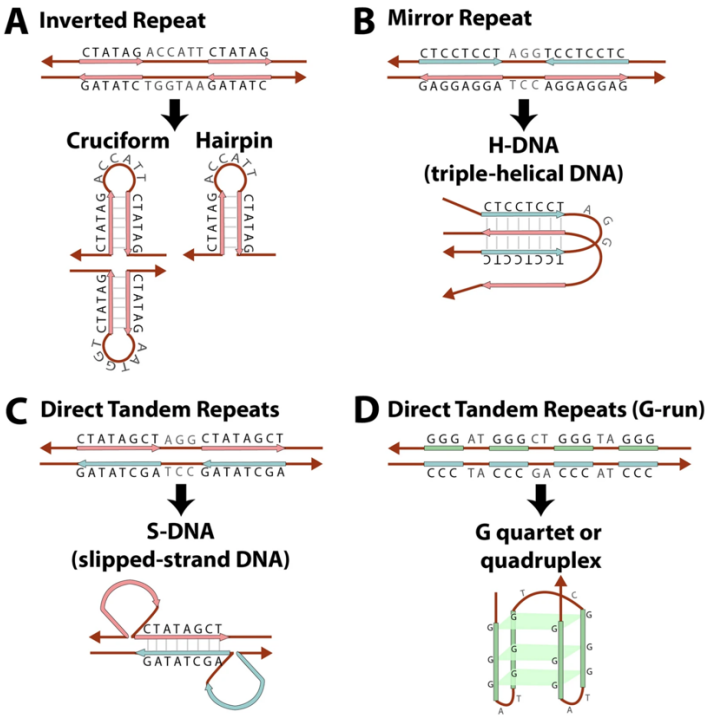


Figure 3. **DNA structures predisposing to fork stalling/breakage and potentially to architectural rearrangements**. **A**. Inverted Repeat can generate either cruciform structures or harpins in double and single-stranded DNA respectively. **B**. Equidistantly identical bases, Mirror Repeat, forms H-DNA or triple-helical structures. **C**. Direct Tandem Repeats (simple, noninterrupted repeats) form S-DNA (slipped-stranded DNA). **D**. Direct Tandem Repeats with G-runs (series of repetitive guanine nucleotides) form G quartet or quadruplex. Figure from (Burssed et al. 2022) open access under Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The role of Low Copy Repeats

The main processes responsible for structural aberrations are homologous recombination (HR) and non-homologous recombination, while the first requires a certain length of homology between sequences (usually 300bp in humans) and proteins to operate, the second can act with none or only microhomology (Reiter et al. 1998; Hastings, Lupski, et al. 2009). The most discussed feature that facilitate the formation of structural variation by HR are Low Copy Repeats, also known as segmental duplications (LCRs, SDs), portions of 100-400 kb sharing up to 97% of sequence similarity to each

other (Stankiewicz and Lupski 2002; Bailey and Eichler 2006). Within two LCRs there is an increased probability of SV formation, indeed, genome regions localized between LCRs tend to form recurrent rearrangements that can have similar breakpoints (within LCRs sequences) and sizes among individuals (Harel and Lupski 2018). The opposite scenario, non-recurrent rearrangements, happens when structural changes do not consistently share these features among individuals, signalling the occurrence of independent mutational events that created them (Figure 4).

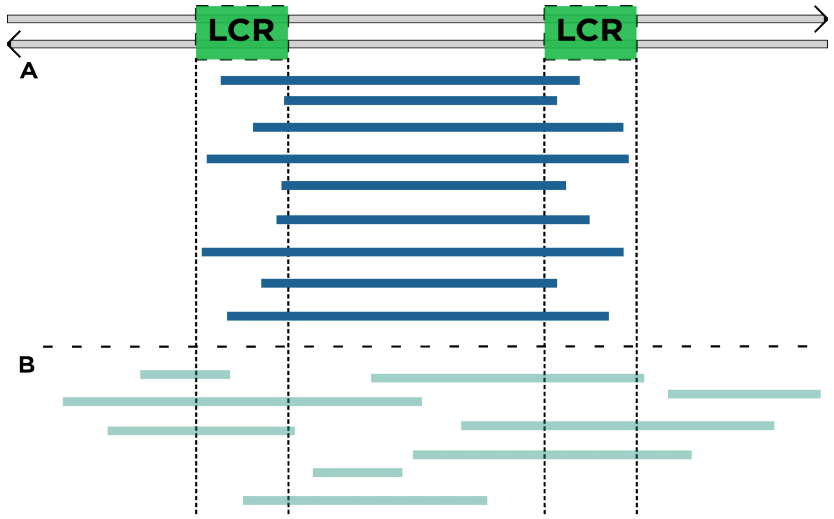


Figure 4. DNA region with two Low Copy Repeats showing recurrent and non-recurrent events. This simplified figure shows putative rearrangements, represented with the coloured bars, in multiple individuals mapped against a general reference **A**. Recurrent rearrangements form within two LCRs, their breakpoints all cluster within each region encompassing the LCR represented with the vertical dashed lines. **B**. Non-recurrent rearrangements have independent breakpoints and sizes to one another, not clustering within LCR regions.

In the human genome the main mechanisms responsible for the formation of structural chromosomal rearrangements are: Non-allelic homologous recombination (NAHR), Non-homologous end-joining (NHEJ), Fork Stalling and Template Switching (FoSTeS) and Break-Induced Repair (BIR).

NAHR

Non-allelic homologous recombination (NAHR) occurs, as the name suggests, during recombination of chromosomes in mitosis or meiosis when two non-allelic LCRs align due to their sequence identity (Gu et al. 2008; Burssed et al. 2022) (Figure 5). NHAR occurring within two LCRs having the same orientation within the same chromosome or between two sister chromatids, causes a deletion and a duplication respectively, when misalignment occurs within the same set of LCRs an inversion is generated and, lastly, a translocation takes place if the event is between different chromosomes (Burssed et al. 2022). Even though this process can happen in mitosis and meiosis, in the latter process NAHR can also occur within one of the two sister chromatids; in this specific case, only deletions can arise from this event. Consequently, this affects the frequency of reciprocal deletions and duplications produced between meiosis and mitosis, which in meiosis is not equal and can be higher for deletions, reflecting the rate at which intrachromatid rearrangements occur (Gu et al. 2008; Turner et al. 2008). Interesting evidence highlighted that LCR sequences have the tendency to cluster in specific locations in the genome and that in their surrounding regions is more likely to find DSBs (Reiter et al. 1996; Bi et al. 2003; Wells 2007). Overall, LCRs not only predispose to chromosomal rearrangements through NAHR, but reside themselves in specific hotspots in regions that, in turn, influence their formation.

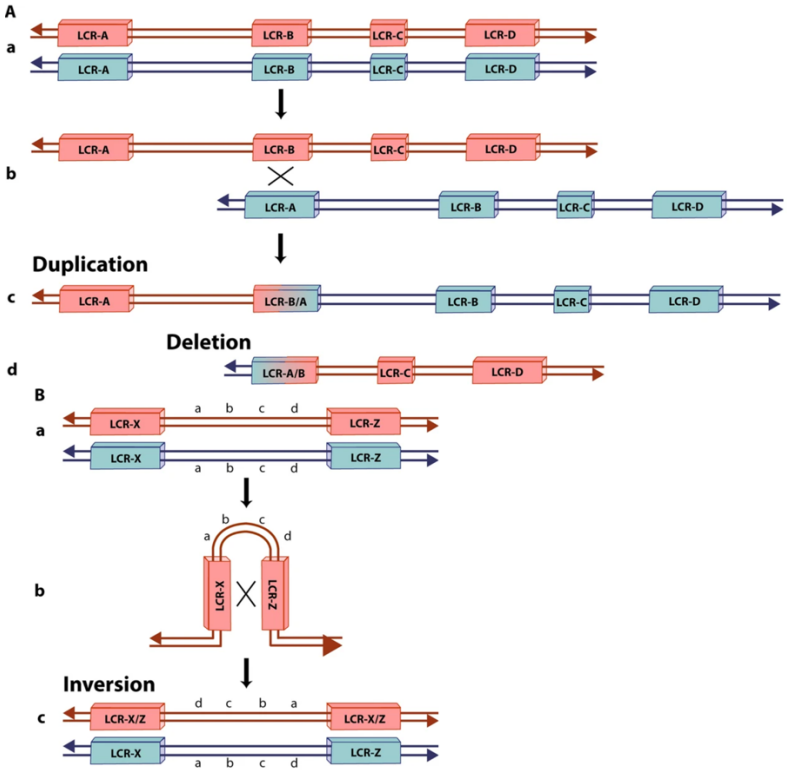


Figure 5. **The mechanism of Non-Allelic Homologous Recombination (NAHR)**. **A**. LCRs mediate NAHR leading to the formation of duplication and deletion: **a**. a standard chromosome pairing and correct allelic alignment of LCRs; **b.** misalignment between LCRs mediated by their high level of identity, which has two outcomes, **c**. a duplication and **d.** a deletion. **B**. NAHR leading to the formation of inversions: (a) correct alignment of LCRs, LCR-X and LCR-Z have similar sequences but arranged in opposite orientations, **b**. in case of misalignment between such LCRs, unequal crossing over generates **c**. an inversion. Figure from (Burssed et al. 2022) open access under Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

NHEJ

Non-homologous end-joining (NHEJ) is the process by which DSBs are repaired after a damage and the canonical mechanism works to re-unite the two broken double strands, usually by connecting, modifying and joining the ends together (Figure 6). Sometimes NHEJ do not alter the base pair sequence interested by the break but can also be imprecise and leave behind repair signs such as gains or losses of nucleotides between the two points of connection, but is also responsible for the formation of duplications, deletions, inversions and translocations (Lieber 2010; Ottaviani et al. 2014; Hattori and Fukami 2020). Although NHEJ can use sequence microhomology to operate, it is not required for its functioning; nonetheless, a specific type of end joining, Microhomology-Mediated End-Joining (MMEJ) requires specifically this homology to repair a DSB. MMEJ leaves at the break site two 3' overhang single strand DNA segments, each one on a DNA strand, exposing microhomology stretches that can anneal due to their similarity, while the non-homologous unannealed sequences are deleted and the 5' resected gaps are synthesized (Lieber 2010). As NHEJ, also MMEJ is error prone, since the region around the repaired DSB is deleted, but the two repair mechanisms require different molecular machineries to function and it is probably their availability in the cell that might guide towards one repair type over the other (Lieber 2010; Ottaviani et al. 2014; Liu et al. 2019).

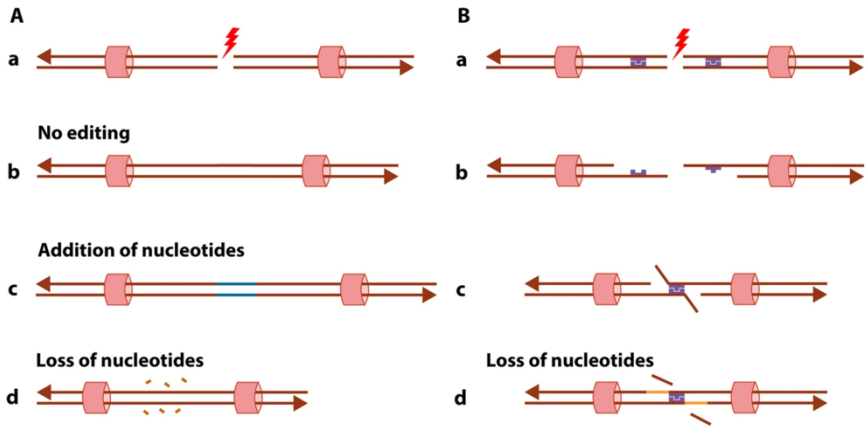


Figure 6. Mechanisms of double-strand break (DSB) repair. **A**. Canonical Non-Homologous End-Joining (c-NHEJ) mechanism: after **a**. the DSB, the molecular mechanism of repair unites back together the broken ends. In this process, nucleotides at the breakpoints can either be **b**. unedited, **c**. added, or **d.** lost. **B**. Microhomology-Mediated End-Joining (MMEJ): after **a**. the DSB, **b.** a 5′ to 3′ resection produces two 3′ single-stranded overhangs exposing microhomology segments (purple), which **c.** anneal, and the **d**. repair produces a deletion. Figure from (Burssed et al. 2022) open access under Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

FoSTeS and MMBIR

Some SVs can be complex in their nature, meaning that they carry different rearrangement types in a defined genomic region. Errors during replication have been proposed as the responsible for such complex rearrangements and a model called Fork Stalling and Template Switching (FoSTeS) was suggested to facilitate their formation (Lee et al. 2007b). In this process, during DNA replication, the replication fork stalls, the lagging strand separates from the original template to invade other replication forks in its surroundings and restarting DNA synthesis, the invading strand can either disengage and return to its original location or continue the process multiple times, acquiring "foreign" DNA segments at each step; the process is mediated by microhomology between invading strand and templates (Lee et al. 2007b). If the invasion occurs at a downstream replication fork a deletion is created, upstream invasion results in a duplication, forks in

different chromosomes create translocations and inversions arise because the lagging strand can proceed in both directions (5'-3'or 3'-5') (Lee et al. 2007b; Burssed et al. 2022).

An additional impairment during DNA replication happens when the replication fork breaks, forming a single-ended DSB as described above. In these cases, mechanisms like FoSTeS, Break-Induced Repair and Microhomology-Mediated Break-Induced Repair (BIR and MMBIR respectively) take place (Figure 7). In this case, resection of the 5' end of the broken DNA leaves a single-stranded 3' end that can invade other regions using long homologous segments for BIR or shorter microhomology ones for MMBIR, the invasion of another double-stranded DNA portion establishes a new replication fork where synthesis proceeds (Burssed et al. 2022). As described for FoSTeS, this process can happen multiple times and the direction of the invasion determines the kind of SV formed. Interestingly, neither NHEJ nor MMEJ manage to resolve this error due to the lack of a second end to be annealed by the 3' overhang (Hastings, Ira, et al. 2009; Ottaviani et al. 2014; Burssed et al. 2022).

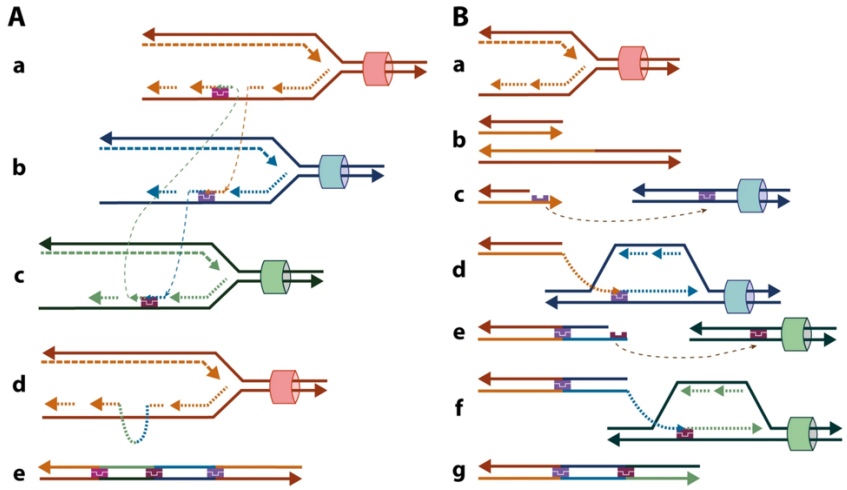


Figure 7. Replication mechanisms. **A**. Fork Stalling and Template Switching (FoSTeS) mechanism: **a**. Stalling of a replication fork, **b**. the lagging strand separates from the template and, it invades (via microhomology, in purple), another template (dashed line) of another active replication fork, restarting DNA synthesis. **c**. This process can be repeated multiple times other replication forks. When returning to its original template and **a**, **d**., the new **e**. product contains segments from different other genome locations. **B**. Microhomology-Mediated Break-Induced Replication (MMBIR) mechanism: **a**. A collapsed replication fork forming **b**. a single-ended DSB. **c**. resection creates a 3′ overhang exposing a microhomology segment (purple), invading **d**. another genome region, where DNA synthesis is restarted. **e**, **f**. this process can occur multiple times and, in the end, **g.** the resulting product is a mixture of distinct parts of the genome rearranged together thanks to microhomology regions. Figure from (Burssed et al. 2022) open access under Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Chromoanagenesis

Besides the aforementioned processes, which can involve portions of DNA sequences along the same or between different chromosomes, also large-scale events, comprising up to the entire length of a chromosome, can take place. Chromoanagenesis generally contains three macroscopically catastrophic events altering the structure of one or more chromosomes (Figure 8) and comprises: chromothripsis, chromoanasynthesis and chromoplexy (Holland and Cleveland 2012). Chromothripsis (meaning chromosome shattering) takes place when multiple DSBs happening along a chromosome cause its collapse and fragmentation, the fragments are then reassembled randomly via NHEJ/MMEJ and different portions might not be incorporated back, leading to loss of genetic material (Ly and Cleveland 2017; Zepeda-Mendoza and Morton 2019; Brás et al. 2020; Hattori and Fukami 2020; Nazaryan-Petersen et al. 2020). In chromoanasynthesis (chromosome reconstruction) a chromosome segment or even the entire sequence undergoes new synthesis utilizing either FoSTeS or MMBIR mechanisms to form newly rearranged DNA segments, this process produces duplications, deletions but also complex rearrangements such as duplications and triplications together (Liu et al. 2011; Weckselblatt and Rudd 2015b; Pellestor 2019; Zepeda-Mendoza and Morton 2019; Burssed et al. 2022). The last large-scale event, chromoplexy, or chromosome restructuring, takes place when more than two chromosomes exchange their parts among each other via NEHJ or MMEJ, forming “chimeric” chromosomes and giving rise to translocation, thus leaving unaltered the genomic dosage, even though marginal nucleotide gain/loss has been reported (Baca et al. 2013; Pellestor 2019; Zepeda-Mendoza and Morton 2019).

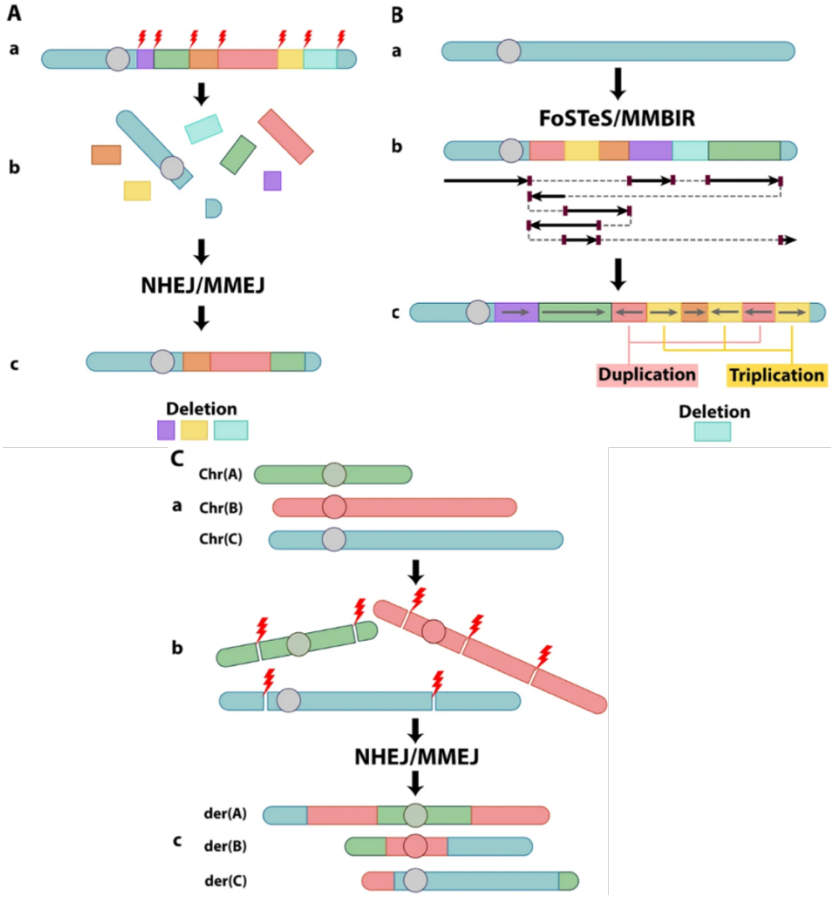


Figure 8. Chromoanagenesis. **A.** Chromothripsis: **a**. A chromosome suffers multiple DSBs, leading its **b.** shattering. **c**. the chromosome is then reassembled by c-NHEJ or MMEJ. Deletions due to loss of DNA fragments can occur. **B**. Chromoanasynthesis: **a**. A chromosome undergoes **b**. either FoSTeS or MMBIR, forming **c**. a newly assembled chromosome, which can present inversions, deletions, duplications, and triplications. **C**. Chromoplexy: **a.** more than two chromosomes are shattered by **b**. DSBs and, after recombination by NHEJ or MMEJ, **c**. form rearranged chromosomes with translocations. Figure from (Burssed et al. 2022) open access under Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The mechanisms leading to the formation of structural rearrangements have also been studied for addressing human genomic disorders and, conversely, investigating health-related SVs provided an excellent boot camp for a deeper characterization of the molecular events causing them. One aspect that links SVs to diseases concerns the direct influence that rearrangements can have on a DNA sequence, in particular their diversity. When recombination events happen between homologous chromatids, for example by the BIR mechanism, they can lead to loss of heterozygosity when the chromatids bearing the same copied allele segregate together and, additionally, in some cases this can result in diseases when the copied stretch either have a pathogenic recessive variant or an epigenetically imprinted locus (Hastings, Lupski, et al. 2009; Carvalho and Lupski 2016). Regions of the genome more prone to rearrangements can be scouted for their higher probability of causing disorders, this is the case of LCR-causing NAHR reciprocal duplication/deletion responsible for Charcot–Marie–Tooth disease type 1A (CMT1A) and hereditary neuropathy with liability to pressure palsies (HNPP) in chromosome 17 (Stankiewicz and Lupski 2002; Lindsay et al. 2006; Carvalho and Lupski 2016). Other syndromes have been connected to NAHR events, such as the DiGeorge, the Williams–Beuren and the Prader–Willi syndromes[2], as well as specific NAHRs mediated by LINE-1 or Alu elements, in line with previous reports

[2] Charcot-Marie Tooth disease type 1A is a neuropathy affecting the signal transmission from the brain to the peripheral areas and is usually characterized by muscle weakness (hypotonia), sensory decay.
Hereditary neuropathy with pressure palsies also affect nerve signal transmission, causing hypotonia, numbness, tingling, and can be accompanied by limb pain and loss of sensation in the hands. DiGeorge Syndrome, caused by the 22q11.2 deletion is associated with heart and behaviour/learning defects, as well as hearing and speech problems.
Williams–Beuren also manifests with heart anomalies and dysmorphic face features and is caused by the 7q11.23 deletion.
A deletion at 15q11.2 causes the Prader–Willi syndrome, which is characterized by hypotonia, short stature and aggressive behaviours.

highlighting how LCRs lead to genomic instability through NHAR events, and loci associated with them in meiosis showed evidence of genomic disorders via deletions and duplications (Lehrman et al. 1985; Shaikh et al. 2000; Stankiewicz and Lupski 2002; Weise et al. 2012; Kohmoto et al. 2017). Intriguingly, as previously discussed in the chapter describing CNV implications in diseases, reciprocal duplications and deletions produced by NAHR can have directly opposed effects on the phenotype they influence. Examples of these actions from reciprocal duplications/deletions reflect on traits as the size of the head, disorders such as schizophrenia and autism and also on weight (Golzio and Katsanis 2013; Carvalho and Lupski 2016), with the first and last example reflecting the effects on their respective phenotypes, microcephaly/macrocephaly and overweight/underweight for duplications and deletions respectively. NAHR is not the only mechanism with studied underlying pathogenic SVs, also NEHJ, BIR and chromothripsis examples exist and underline the importance of knowing and understanding such molecular dynamics in order to predict additional loci that could putatively predispose to diseases.

1.1.3 Detecting structural variants

Since the early approaches screening individual karyotypes to detect uncommon macroscopic rearrangements of genomic portions in the 20th century, researchers looked at increasingly refined and precise methodologies to better identify structural variations in the studied genomes, as the technological advancements progressively allowed so. The first major concern was to be able to go deeper at the genome level and to detect previously invisible variants, which were escaping more rudimental discovery techniques simply due to their intrinsic limitations in size accuracy. More recently, instead, the focus has shifted towards the integration and refinement of different technologies for detection of SVs and a higher power to resolve more complex events at the same locus, rather than single occurrences.

Wet lab approaches

One of the first laboratory techniques that allowed the detection of large genomic rearrangements is chromosome banding, namely the procedure by which chromosomes in their condensed form are stained during the S phase[3] of the cell. Even though there are different banding methods like Giemsa, reverse, quinacrine or centromere that stain at different intensities chromosomal regions, the most used methods are primarily Giemsa (G) and reverse (R), with the ability to differently stain heterochromatin (A/T-rich regions) and euchromatin (G/C-rich regions) regions (Caspersson et al. 1968; Francke 1994; Trask 2002; Balachandran and Beck 2020). Banding not only allows the identification of chromosomes and their intrinsic structure (karyotyping) but also the comparisons of banding patterns among

[3] The classic stages of cells' life can be divided into an interphase, where the cell is not macroscopically changing, and division, where the cell divides into to the two daughter cells. Within interphase, cells organize the subsequent mitosis by preparing DNA replication machinery (G1 phase), actively synthesizing new DNA (S phase) and getting ready for cell division (G2 phase).

different cells to infer differences in terms of deletions, insertions and, particularly, translocations; nonetheless, due to its intrinsic low resolution, only large (>3Mb) rearrangements can be detected (Balachandran and Beck 2020).

A later and more refined technique for localizing specific sequences in their native environment, fluorescence in situ hybridization (FISH), relies on fluorescent probes that ligate to target DNA sequences and can be subsequently analysed to assess the localization of probes (Bauman et al. 1980). For SV detection, FISH allowed a much higher resolution (with refined techniques reaching ~1kb) while providing lower false positive rates than banding, while detecting the same set of rearrangements and better reach more difficult localizations, like sub-centromeric regions (Pinkel et al. 1988; Kallioniemi et al. 1992; Linardopoulou et al. 2005; MacKinnon and Campbell 2013; Cui et al. 2016).

SNP array detection and optical mapping

Another technique to infer SVs, and specifically CNVs, was developed using SNP arrays as a proxy to infer the presence of such structural rearrangements. SNP arrays, widely used to genotype thousands of single nucleotide variants in samples and consequently also used in population genetics, are biochips designed as arrays of immobilised oligonucleotides bearing a specific allele of the studied SNPs panel. These allele-specific oligonucleotides exist in two forms, each carrying one of the two alleles of the investigated SNP, that are annealed with the sample DNA sequences carrying those alleles. Depending on the technology used, either by perfect match with fluorescently labelled sample sequences or single-base extension of the tested allele with the incorporation of a fluorescent labelled nucleotide, colour-coded signals for the four DNA bases are emitted and analysed at each binding cycle (LaFramboise 2009). While the first algorithm to detect CNVs from SNP array data was based on a study of

cancer genomes (Zhao et al. 2004), other researchers utilised more straightforward methods to begin detecting specifically deletions, without the need of mining information with custom algorithms. These methods relied on the absence of calls and on violation of Mendelian inheritance. Briefly, the first inference method is based on the fact that samples having a homozygous deletion at a locus, simply cannot be called for those SNPs residing within the deleted regions resulting in a "No Calls"; the latter method instead analyses mother-father-child trios and detects inconsistencies with the expected AA, AB or BB alleles, due to inherited deletions (Conrad et al. 2005; McCarroll et al. 2006; LaFramboise 2009). Specific algorithms fully analysing SNP array data consider two sources of information. The first one estimates the copy number of a SNP by summarizing probe intensities and comparing it to a panel of standard samples, the second method utilizes the B allele frequency (BAF, namely the frequency of minor allele) information to analyse the observed BAFs, that in case of gains or losses do not match the expected frequencies (Zhao et al. 2004; Lai et al. 2005; Wang et al. 2007; LaFramboise 2009). As reported above, the detection of SVs via SNP arrays is only limited to gains and losses, while mutations such as inversions or translocations cannot be detected, preventing a complete characterization of structural rearrangements (Wang et al. 2007; Balachandran and Beck 2020).

One last fluorescent-based technique to infer architectural alterations in genomes is high-throughput optical mapping. It utilizes fluorescent labelling, similar to FISH, to pinpoint the action of restriction enzymes nicking stretches of DNA (300 kb-3 Mb in range) to be optically imaged to study read information (Teague et al. 2010). The reads produced with this restriction experiment are *de novo* assembled into contigs and compared against a reference to infer rearrangements, this can be also done to yield high-throughput results using Bionano platform to discover deletions (>

500 bp), duplications (> 30 kb), insertions (> 500 bp), inversions (30 kb) and translocations (> 50 kb), with the drawback to miss a breakpoint resolution variant (Chan et al. 2018; Balachandran and Beck 2020). Overall hybridization techniques are advantageous because of their low cost at processing hundreds of samples, but they only detect CNVs and do not resolve them at a base-pair level. For these reasons, SV analysis, and more broadly genome-processing technologies, moved towards the implementation of whole genome sequencing to obtain a more fine-grained information (Balachandran and Beck 2020).

Whole genome sequencing detection

With the advent of high-throughput whole genome sequencing (WGS) a new way to efficiently explore the genome of many samples had opened, continuous technological advancements provided increasingly higher output capacities and produced a drastic reduction of processing costs (Kircher and Kelso 2010; Wetterstrand 2013; Pevzner and Compeau 2016). Even though during the history of genetic research a special place has always been reserved for SNVs, mainly due to the practical investigation reasons and a well-oiled analysis mechanism, by means of the aforementioned detection techniques also SVs started to attract attention and these new research possibilities paved by Next Generation Sequencing (NGS) sparked numerous studies addressing structural rearrangements. Many studies, indeed, uncovered how SVs are widespread in the human genome and how they can have direct effects over regulation of biological processes, transcription and genome organization (Weischenfeldt et al. 2013; Sudmant, Rausch, et al. 2015; Spielmann et al. 2018). It was rapidly noticed how their newly discovered importance made them particularly important when it is needed to assess on the one hand the impact of these variants in normal and altered health statuses and, on the other hand, their

role in evolutionary processes and dynamics in human population genomics.

The first available technology in NGS, often still widely used nowadays is short-read sequencing. In this sequencing type, after the sample DNA has been sheared into smaller fragments, both fragment ends of a specifically predefined size are sequenced in what is called paired-end sequencing – as opposed to single-end where sequencing only occurs at one of the ends of the fragment. The reads produced this way are then mapped against a reference genome to identify their original location. Even before SV calling, the mapping phase itself is still not an optimal step. This happens because the reference genome used in research still contains long stretches of repetitive sequences and gaps that hinder or completely prevent the mapping step - in particular around centromeres and telomeres (Balachandran and Beck 2020). Notwithstanding these difficulties, mapping still performs well in many positions and it has been noted that the way how reads map to the reference is a relevant and valuable information that can be leveraged to infer the presence of SVs. Different approaches emerged to exploit the signature information left by mapped reads and, during the last decade or so, many different software have been developed to detect SVs from short-read sequence data, the vast majority initially only adopted a single detection method (Alkan et al. 2011). Nonetheless, researchers did not wait too much until multi-detection approach software started to be publicly available and the older single-strategy methodologies were sometimes used less in favour of these last improved algorithms. The different detection methods are: read-depth, split-read, read-pair and *de novo* assembly (Figure 9). These methods to leverage read information rely on different signatures occurring during the mapping phase. The read-depth method assumes that the coverage of a region relates to the number of copies of that region and scans genome portions to evaluate the density of

reads mapping at those locations (Alkan et al. 2009; Yoon et al. 2009; Abyzov et al. 2011a). To assess this, the method counts the number of reads mapping to fixed size regions (bins) and, after data normalization, estimates absolute copy number for each region (Sudmant et al. 2010; Tattini et al. 2015). Although the method is good at estimating the copy number of a region, and thus infer deletions and duplications, the resolution of the breakpoint of the event is poorer than other methods (Escaramís et al. 2015). The split-read approach takes advantage of how paired-end reads map: if one of the paired reads accurately maps to the reference while the other does not or only maps partially, the latter may signal that this latter read spans the breakpoint of an SV. The splitting of the unmapped (or partially mapped) read in sub-reads allows for a second mapping step of these portions independently. After this step, the two portions of the split read will flank respectively the start and end point of the detected SV and thus, for its intrinsic nature, this method provides base-pair resolution, which might nonetheless vary due to microhomology at breakpoints (Zhao et al. 2013; Schröder et al. 2014; Escaramís et al. 2015). In the read-pair approach, SVs are detected by relying on the spacing between read pairs mapped to reference (Korbel, Urban, Grubert, et al. 2007; Korbel, Urban, Affourtit, et al. 2007). Pairs of reads mapping closer or further to one another than what is expected based on their average insert size, signal deletions or insertions. Inverted read orientation is a sign of inversions and lastly mapping on different chromosomes highlights translocations (Zhao et al. 2013; Escaramís et al. 2015). The last method, *de novo* assembly, uses groups of overlapping reads to create contigs; these sequences are longer than short reads and represent the union of non-repetitive information that the latter provide. The comparisons between contigs and a reference genome highlight regions with putative discordant copy number, where SVs may occur, thus potentially discovering novel insertions (Hajirasouliha et al. 2010; Mills et al. 2011; Tattini et al. 2015). Even though the assembly-

based method can be prone to error during the assembly build-up and has a high computational cost, it can be a good tool to refine the outcome of other approaches at targeted regions (Escaramís et al. 2015). It is also worth mentioning that these contigs can be used to exploit "leftover" reads that could not map efficiently in the aforementioned complex genome regions, and that, in a conventional mapping against a reference, would be normally discarded (Nagarajan and Pop 2013; Chaisson et al. 2015).

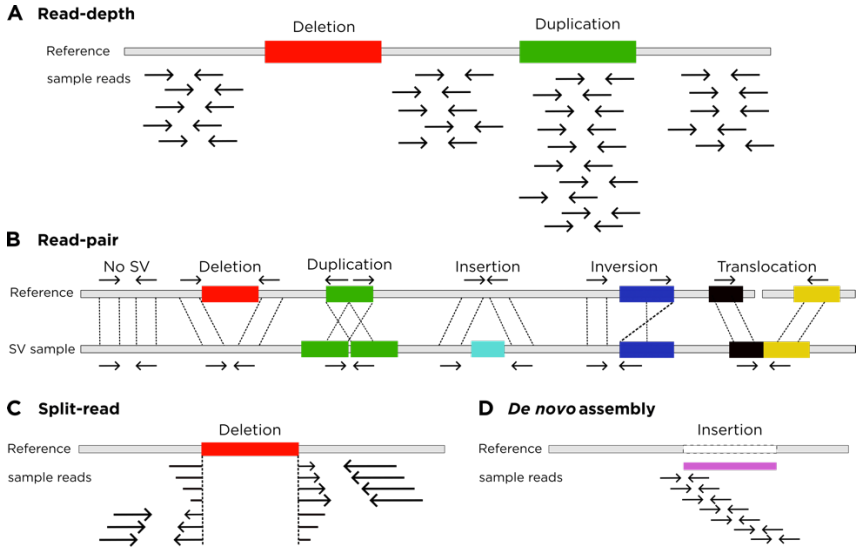


Figure 9. Detection methods for SVs. A. Read-depth evaluates difference of coverage of mapping reads, deleted regions in the sample will produce detectable decreases of mapping reads in that region, conversely genome portions with significantly higher amount of reads signal the presence of duplications in the sample. **B**. Read-pair. The mapping pattern of read pairs is evaluated, from left to right: 1) no SV present, reads map as expected; 2) a deletion is detected when the aligned pairs map further apart than what expected based on the insert size; 3) in tandem duplications, read pairs are align in an unexpected order, highlighting that the genome portion they reside has been copied and relocated; 4) insertions, are detected because reads are aligned closer than expected based on the insert size; 5) reverse read orientation is indicative of an inversion event, both reads align either in forward or reverse strand; 6) translocations are detected by read pairs mapped to different chromosomes. **C**. Split-read. Reads spanning the breakpoint of an SV are split at the breakpoint site when mapped to the reference, while the mate paired read is properly aligned. **D**. De novo assembly. An insertion is identified by the assembly of sample reads into a contig that is not present in the reference genome. Figure adapted from Escaramís et al. (2015)

Limitations and new approaches

Even if all these different approaches have been developed, it is important to make clear that the software implementing these algorithms available nowadays, still struggle to overcome the intrinsic limitations (insert size length) of short-read WGS and hence, due to differing detection power and to the extent of SV types and length, it is not possible to retrieve a complete picture of structural variation within a genome using a single caller (Telenti et al. 2016; Chaisson et al. 2019; Kosugi et al. 2019; Lappalainen et al. 2019). It should be stressed that the software developed so far are well established and, far from being a compromise in research, they are one of the optimal choices available to computationally infer structural variation from short-read data. With that said, unfortunately it is true that on average the recall lays between 10 and 70% while the false positive rate is still very high (up to 89%), depending on the type of SV and its size (Mills et al. 2011; Teo et al. 2012; English et al. 2014; Sudmant, Rausch, et al. 2015; Tattini et al. 2015; Huddleston et al. 2017a; Jeffares et al. 2017; Sedlazeck, Rescheneder, et al. 2018a). With this in mind, still, the majority of the initial small and large-scale research on SVs essentially relied on these methods with an overall satisfactory precision at the expenses of recall[4] (Sudmant, Rausch, et al. 2015; Sedlazeck, Rescheneder, et al. 2018a; Audano et al. 2019; Mahmoud et al. 2019). One of the main strategies to mitigate the low

[4] Precision and recall are the metrics expressing the performance of a classification algorithm performing a labelling task. More specifically, to assess the performance of different algorithms for SVs detection, their action is tested against a sample with known results (all the SVs present in the sample, in this case). Precision is the proportion of relevant elements among all retrieved elements. If a sample has 100 known SVs, the maximum number of relevant (real) elements that the software can detect matches this quantity, while any other SV (spurious) outside this 100-element set will just be a false positive. If a caller detects 80 SVs but only 60 are within the true set, then it recovers 60 true positives and 20 false positives, its precision, then, will be 60/80 = 0.75. Recall instead expresses the fraction of true elements over the set of all the original elements retrieved. Using the case above, the recall would be 60 retrieved true SVs out of 100, 60/100 = 0.6.

recall of many software consists of using multiple callers to leverage together the strength of different algorithms to build up an integrated dataset, which proved to produce concordant SV calls compared to reference sets (Wong et al. 2010; Lam et al. 2012; Parikh et al. 2016). Although union of algorithms helps in obtaining more accurate results, one downside of this approach is that different studies adopt different combinations of algorithms to build up their own sets of variants, resulting in a tailored rather than standardized approach, thus limiting reproducibility (Ho et al. 2019). Furthermore, often researchers carrying out different projects also use *ad-hoc* methodologies to create the final consensus set of SVs found in their studies. Coordinates overlap, variant distances and false discovery rate (FDR) thresholds, among others, concur to a different degree to how projects integrate raw data – which can be the result of sets of three up to 19 algorithm combinations (Mills et al. 2011; Sudmant, Rausch, et al. 2015; Hehir-Kwa et al. 2016a; Werling et al. 2018; Kosugi et al. 2019; Abel et al. 2020; Collins et al. 2020). These factors together fail to produce a standardized calling pipeline across studies and even if the dataset construction proves to be homogeneous among projects, when those same investigations make use of short-read data, as mentioned above, it remains difficult to overcome intrinsic limitations of the technology.

Computational tools handle the useful marks that SVs leave in the genome while mapping reads to a reference, and short-read technology presents sub-optimal features that hinder a more complete rearrangement scenario. Nonetheless, also sequencing errors could blur the marks of variants and since SVs may cover large portions or even be larger than reads, this generally complicates mapping (Sedlazeck, Rescheneder, et al. 2018a). In addition, it can be difficult to discern types of SVs (tandem duplications or novel insertions) and events can also overlap or reside within a complex rearrangement region, making it harder to disentangle the different mapping

patterns (Sanchis-Juan et al. 2018; Sedlazeck, Rescheneder, et al. 2018a). To ameliorate the features for which short-reads sequencing does not excel, long-read technologies emerged as methods to cover larger portions of the genome with a set of reads. Two main long-read sequencing technologies are the most used in research, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio single molecule real time sequencing (SMRT) uses the advantages offered by an immobilized polymerase on the surface of a nano well, through which single-stranded DNA passes to produce long reads (thousands of bp long, typically more than 5kb) that greatly help read mapping with higher confidence and better capture large SVs than short-reads do (Chaisson and Tesler 2012; Lee and Schatz 2012; Chaisson et al. 2014; Sedlazeck, Rescheneder, et al. 2018a; Audano et al. 2019). Even though PacBio sequencing is very finely tuned at detecting small insertions within mobile elements and small-medium sized variants, one drawback is the high error rate compared to short-reads (Chaisson et al. 2019; Ho et al. 2019). The high per-base error rate is not the only limitation, at the same coverage than short reads, indeed, the sequencing cost is higher, restricting the analysis to fewer samples with the same budget or lower coverage at the expense of shorter read length, thus potentially influencing the study design (Rhoads and Au 2015; Goodwin et al. 2016). Nanopore sequencing (ONT) also makes use of the passage of single stranded-DNA through a molecule, in this case a fixed protein pore. During the passage of DNA, each nucleotide produces a specific change in the electric current of the protein, which is recorded and associated to one of the four possibilities, a proxy-detection similar to what happens with fluorescent detection of incorporated bases in Illumina platforms (Clarke et al. 2009; Eid et al. 2009). Being a long-read technology, also ONT is error prone compared to PacBio, but can be more cost effective reaching higher throughput (Ho et al. 2019). The main signature used to infer SVs from long-read sequencing are split-read and soft clipped read, the event is

reconstructed by either clustering or de novo assembly and, as for short-reads, different software have been devised for this task (Balachandran and Beck 2020). Using long-read approaches, generally developed to overcome the shortcomings from the use of shorter reads, provide a valuable tool to detect with higher precision and recall many SVs and particularly new insertions (Chaisson et al. 2019; Ho et al. 2019; Leija-Salazar et al. 2019; Wenger et al. 2019). The advantages of this technology unfortunately come with shortcomings, mainly the high (5-15%) per-base error rate, which, depending on the technology (PacBio or ONT) is more likely to happen in specific DNA motifs such as 1bp indels or homopolymeric nucleotides (Jain et al. 2018; Sedlazeck, Lee, et al. 2018; Balachandran and Beck 2020). All things considered, great advancements have been made in SVs detection thanks to these newer technologies and research is constantly pushing advancements forward to refine these approaches and obtain cleaner data at a cheaper price. To prove the undeniable progress, it was estimated that long-read sequencing improve the detection of SVs by three-fold compared to short-reads and excels in the identification of long transposon insertions as well as SVs in repetitive regions (Chaisson et al. 2019).

Other methods to gather information at genome level include Linked-Reads and Hi-C. In Linked-Reads (10X Genomics Linked-Reads, LR), the spatial information of molecules is conserved via a molecular barcoding so that the original location of each short-read fragment can be traced back in order to reconstruct the long-range interactions among reads (Zheng et al. 2016). The features of linked-reads sequencing make it a suitable method for large CNVs and translocations (Zheng et al. 2016). The high throughput chromosome conformation capture (Hi-C) is capable of sequencing DNA and retaining the information of sequences mapping close together in the overall 3D space, that might not be close in the linear conformation (Lieberman-Aiden et al. 2009). As Linked-Reads, the Hi-C sequencing is

suited for large rearrangements and translocation detection, as it can discover read-pairs spanning megabases apart in the genome whose special configuration brings the sequences close to each other (Ho et al. 2019).

In general, nowadays current methodologies for the discovering of SVs rely on a multiplatform approach, where new technologies are useful to detect novel variants and resolve complex scenarios, while short reads are used to account for the recurrent errors of long reads, in a process known as polishing (Sedlazeck, Lee, et al. 2018; Fu et al. 2019; Ho et al. 2019; Lima et al. 2020; Zhang et al. 2020; Zook et al. 2020). As an example, a recent study on three different trios using nine platforms, discovered an average of almost 30,000 SVs per individual, marking a strong difference with the ~12 variants estimated when SVs detection started (Iafrate et al. 2004b; Chaisson et al. 2019).

1.1.4 Implications in evolution

As for the evolutionary implications of CNVs, many factors come into play when evaluating the function of variants and the selective forces acting on them. As mentioned above, SVs in general act in two dimensions, spanning a fragment of DNA as opposed to SNPs, thus intrinsically operating on more genome content than point mutations do, and theoretically making it more likely to alter normal sequence functions. Variant localization also contributes to the effects of structural alterations, while purifying selection usually acts on CNVs overlapping genes. Cases of positive selection have also been proposed as the driving forces for these variants and, in general, genic CNVs lacking deleterious effects are more prone to be adaptive than intergenic mutations (Cooper et al. 2007; Hurles et al. 2008; Conrad et al. 2010).

Positive selection on CNVs

A common example

Arguably one of the most famous examples of influence of copy number variation over adaptive phenotypes, in recent evolutionary human history, is the case of the *AMY1* gene, which encodes an amylase enzyme responsible for the metabolization of starch. This gene is expressed in the pancreas as well as in salivary glands and is present with multiple copy number configurations in humans, resulting in variable copy number both among human populations and more broadly within mammalians (Robyt and French 1967; Hagenbüchle et al. 1980; Bank et al. 1992; Iafrate et al. 2004b; Boehlke et al. 2015; Pajic et al. 2019). Intriguingly, it was pointed out not only that *AMY1* copy number is positively correlated with the levels of salivary amylase protein - hence providing the ability to digest starch more efficiently - but also that individuals from populations traditionally using more starch in their diets had on average more *AMY1* copies than individuals from low-starch diet populations (Perry et al. 2007). Indeed, it

is argued that the expansion of *AMY1* could have helped members of agricultural subsistence societies to better cope with the major nutrient entering in their diets, thus having reinforced metabolic machinery allowing to efficiently metabolize energy-rich molecules, which presumably would have been selected positively during human history. Research on *AMY1* copy number also extended beyond oral cavity capabilities to digest starch and revealed that the number of copies is also related to response to pathogens and microbiota composition, possibly shifting the simpler scenario of dietary aid system, to broader general implications (Pruimboom et al. 2014; Poole et al. 2019). Regarding the extra-oral implications of the copy number variability of *AMY1*, reports show that individuals with fewer copies of the gene retain higher levels of blood glucose after starch loading and are more prone to obesity (Falchi et al. 2014; Higuchi et al. 2020). Despite research strongly indicated the possibility of selection acting on this locus, its actual genomic architecture poses challenges to accurately reconstruct the breadth and timing of the forces that took place; indeed, the multiple copy configurations (2 to 17) and the likely occurrence of events disrupting LD in the locus complicates the investigation even further, making the dynamics of this locus still not fully comprehended (Popadić and Anderson 1995; Usher et al. 2015; Saitou and Gokcumen 2019a).

Brain functions

Positive selection signals have been detected in human specific duplication-driven gene families in comparative genomics analyses within great apes, as well as for CNV-containing genes related to immune response, brain function and reproduction (Johnson et al. 2001a; Dumas et al. 2007; Perry et al. 2008; Han et al. 2009; Gazave et al. 2011; Niu et al. 2011; Iskow et al. 2012). Noteworthy investigations report functional implications of CNVs in both disease risk factors and human evolutionary history. For example, deletions and duplications localized in chromosome 1q21.1, were

found to be correlated with head size in subjects with micro- and macrocephaly, respectively (Brunetti-Pierri et al. 2008). Intriguingly, in humans a copy of the *HYDIN* gene in chromosome 16 - a candidate for causing hydrocephalus - is duplicated almost completely in 1q21.1, and deletions at the sequence original location have been associated with microcephaly (Fujiwara et al. 1992; Callen et al. 1993; Davy 2003; Doggett et al. 2006). Another notorious duplication event in Hominoidea[5] involves the complex duplicated sequences discovered in human chromosome 16 (Stallings et al. 1993; Loftus et al. 1999). Comparisons of duplicated segments among hominids revealed a textbook example of extreme positive selection, where nonsynonymous mutations specifically accumulated within the *morpheus* gene family (Johnson et al. 2001b). Considering the number of alterations and their strong tendency to change the amino acid sequence in this gene family, also known as nuclear pore complex interactive protein (NPIP), it has been suggested that NPIP was subject to adaptive evolution during the evolution of African hominids (Johnson et al. 2001b). Even though the function of the *morpheus* gene family is not well established yet, its products apparently interact with the nuclear pore complex and might have a role in mRNA trafficking across the nucleus (Davis and Blobel 1986; Johnson et al. 2001b; Iskow et al. 2012). Adaptive selection nearby this locus likely continued in *Homo sapiens*, where a ~280 kya[6] old species-specific duplication of *BOLA2* gene emerged, under positive selection, at chromosome 16p.11.2 (Nuttle et al. 2016). Copy number variations at this locus have been reported to associate with autism and schizophrenia (Kumar et al. 2007; Marshall et al. 2008; Weiss et al. 2008; McCarthy et al. 2009; Jacquemont et al. 2011); thus, the human specific *BOLA2* duplication not only influences the transcription and

[5] Primate superfamily comprising Gibbons, Orangutans, Gorillas, Chimpanzees and Humans
[6] Kya: Kilo (1,000) years ago

dosage of this iron homeostasis gene but adds up to the rearrangement events associated to neurodevelopmental conditions (Haunhorst et al. 2013; Banci et al. 2015; Nuttle et al. 2016). Also, conserved human-specific deletions are candidates of adaptive selection: these variants, which are absent in other primates and mammals and tend to reside outside genes, map close to regulatory elements (McLean et al. 2011). Two interesting examples of losses, that occurred exclusively in the human lineage, pose interesting cases of evolution. The first include a deletion removing an enhancer of a brain-expressed growth inhibitory gene (*GADD45G*) and might have facilitated brain expansion during human evolution; the second example concerns a deletion of the androgen receptor gene (*AR*) enhancers, whose non-expression or disrupted form prevent the formation of penile spines, present in different animals (Murakami 1987; Zhang et al. 2002; McLean et al. 2011; Banci et al. 2015; Orr and Brennan 2016). Even though the latter examples could not be formally tested for signatures of selection, the phenotypes associated to their emergence may prompt speculation about an adaptive role in humans, whether it could be the progression towards larger brains within the Homo lineage or changes in copulation and formation of monogamous strategies of reproduction (McLean et al. 2011; Dixson 2012).

Hints of balancing selection

Human copy number differences at α-globin genes account for cases of α-thalassemia incidence in human populations; usually, a set of two copies of α-globin is carried on each homologous chromosome, but CNVs contribute to a varying number of copies of the gene, ranging from 0 to 6 diploid copies (Goossens et al. 1980; Flint et al. 1986). Carrying just two diploid copies of α-globin results in α-thalassemia, while the loss of one copy results in a mild form of anaemia; however, the complete ablation of the four canonical copies is incompatible with life (Flint et al. 1986). The distribution of

malaria has been studied in countries where most cases of the disease in α-globin genes occur and the frequency of each deletion configuration follows a specific frequency cline in Southeast Asia (continental and insular), where both cis and trans heterozygote forms of the deletion associate with milder forms of malaria compared to individuals with the regular four copies of the gene (Flint et al. 1986; Lau et al. 1997; May et al. 2007). In this case, a canonical positive selection scenario would not properly fit with the persistence of the two distinct deletion forms, which instead could be preserved by balancing selection, given the advantage against malaria they provide. Another peculiar example of balancing selection for a health-related trait is the case of *DMBT1* gene. Its protein product, DMBT1, belongs to the scavenger receptor protein family capable of binding to different bacteria and host ligands on mucosal surfaces and possesses an internal repeated scavenger receptor cysteine rich (SRCR) domain for its binding (Mollenhauer et al. 1997; Prakobphol et al. 2000; Loimaranta et al. 2005; Reichhardt et al. 2017). CNVs overlapping the SRCR domains in *DMBT1* have been identified as drivers of diversity in humans, with different copy number alleles resulting in 7 to 21 SRCR domains (Sasaki et al. 2002; Polley et al. 2015). Two different haplotypes in humans, one with lower number of SRCR repeats than the other, may indeed be under balancing selection, as suggested by selection scans of the two regions in different populations (Alharbi et al. 2022). Even though the precise selective mechanisms that brought the two alleles of *DMBT1* to persist among populations are not thoroughly established, the authors offer a few scenarios that may be responsible, such as the classic heterozygote advantage[7] or environmental factors across regional groups favouring specific alleles over others. Even though there is no strong evidence of

[7] Heterozygote advantage is the condition where the fitness associated to a heterozygote genotype is higher than that of either homozygote and hence such allele combination is favoured by selection.

relationship between *DMBT1* and specific disorders, its expression profile (amniotic fluid, lungs, intestine, saliva) and the prevalent presence of protein in the earliest stool in new-borns make it tempting to speculate an interaction with human microbiome and a relevance for immune processes (Alharbi et al. 2022).

1.1.5 Implications in health

Although the discovery of worldwide population variability for CNVs brought an unprecedented appreciation for this relatively novel marker, researchers focussed their attention also towards the functional consequences of copy number alterations in humans and other organisms. Case-control studies set up to establish the extent of the SV impact over a specific trait started shedding some new light upon the involvement of this variant class in human health. Though the first assays relied primarily on wet-lab approaches, usually limiting the detectable information to larger-sized events, developing technologies allowed further advancements in analytical techniques and increasingly finer resolution and genome coverage per assay. Consequent to the explosion of cost-effective whole genome sequencing data and the development of advanced strategies, more samples could be analysed simultaneously at a deeper level, making possible to analyse large sets of individuals to better characterize the trait being studied. In the following part, a description of the most common examples of health related CNVs in humans is provided, summarizing the knowledge accumulated to date and exposing the importance of research in the field of copy number alterations in the biomedical studies.

Drug response

Noteworthy examples of SVs implication in health concern genes related to drug response in humans. For example, *CYP2D6* is a known case of CNV-related drug metabolism. Cytochrome P450, or CYP, is a large protein family found in many prokaryote and eukaryote species, and humans carry almost 60 functional sequences and as many pseudogenes (Munro and Lindsay 1996; Zanger and Schwab 2013; Kawashima and Satta 2014). *CYP2D6* is among the few functional genes whose role is related to the metabolism of most common drugs (Zanger and Schwab 2013). It is copy number variable in humans and other primates, and the gene family

may have expanded - acquiring new functions - to protect against different plant toxins (Heim and Meyer 1992; Yasukochi and Satta 2011). Reports show that metabolization rates of an antihypertensive drug in people carrying multiple copies of *CYP2D6* turned out to be higher than wild-type carriers having two copies, while low copy number of the gene leads to hypersensitivity to painkillers (Johansson et al. 1993; Elkalioubie et al. 2011). One hypothesis is that selective pressures on *CYP2D6* varied through time, first being more stringent to better cope with environmental toxins while subsequently relaxing when humans could better control their food intake and interaction with the environment (i.e., agriculture and farming as opposed to hunter-gatherer lifestyle) (Kimura et al. 1989). Regardless of the different evidence accumulated about the selective forces acting on CNVs, the extent of copy number variation in the human genome probably does not fit in a simple model of positive and purifying selective forces shaping human CNV landscape based on their effect on individual fitness. As shown promptly in the early research on copy number differences in humans, these mutations were found in phenotypically normal individuals and contributed to human variation (Sebat et al. 2004), suggesting that CNVs, at least in part, probably evolved under neutrality and a portion of the extant diversity was shaped by drift, demographic events and mutation rates (Iskow et al. 2012).

HIV

One case of copy number variation, influencing a notorious health trait, is the *CCL3L1* gene and its reported influence over HIV infection due to dosage effects. The CC chemokine ligand 3 like-1 (*CCL3L1*) gene encodes a ligand for the CCR5 (CC chemokine receptor type 5) receptor, which is also used as a coreceptor by HIV to enter target cells. The abundance of the *CCL3L1* product is correlated to its ability in competing with HIV for their common CCR5 target receptor. It was proved in studies addressing

individuals in worldwide distributed populations that lower copy numbers of *CCL3L1* are associated with greater susceptibility to HIV infection than the higher copy number allele (Gonzalez et al. 2005; Liu et al. 2010; Mohamad Isa et al. 2020). Another example of interaction between copy number variation and HIV involves the killer cell immunoglobulin-like receptors (KIR). These receptors are usually expressed in some lymphocyte types, such as natural killer (NK) cells, which control viral infection via interaction with major histocompatibility complex (MHC) molecules on the cell surface of target infected cells. Interestingly, studies showed that two allelic variants of a gene *KIR3DL1* and *KIR3DS1*, found at the same locus, may have an influence over HIV infection (Qi et al. 2006; Alter et al. 2007; Alter et al. 2011). Even though this potential controlling action over HIV infection could not be confirmed (Gaudieri et al. 2005; Barbour et al. 2007; Long et al. 2008), intriguingly this polymorphic locus could present both *KIR3DS1* and *KIR3DL1* on the same chromosome, suggesting the presence of a CNV at this location (Martin et al. 2003; Williams et al. 2003). A study on more than 2000 HIV-affected individuals of European ancestry highlighted how this KIR locus is within a copy number variable region, the number of *KIR3DS1* copies is inversely correlated with HIV viral load, and in the presence of *KIR3DS1* and *KIR3DL1* copies constrain more vigorously HIV-1 replication (Pelak et al. 2011).

Cancer

Notwithstanding the difficulty of research with SVs in general, medical investigation is increasingly highlighting the role of these variants in cancer, paving the way to ameliorate early prognosis, therapy response and better elucidate tumour progression. One main issue of short read sequencing still concerns that a part of the genome remains inaccessible to reads, mainly due to repetitive regions and, oddly enough, those regions are usually more prone to form SVs (Sudmant, Rausch, et al. 2015; Carvalho

and Lupski 2016; Zhao et al. 2021). This fundamental fact about the way SVs are discovered consequently leads to the possibility that the estimated contribution of these markers in tumour progression has been misjudged; nevertheless, newer long-read sequencing technologies allow discovering novel SVs, making up for most of those missed by conventional short-read approaches (Sedlazeck, Rescheneder, et al. 2018b; Zhao et al. 2021; Hamdan and Ewing 2022). As for the implications of SVs in cancers, when comparing healthy and tumour genomes, the latter show peculiar types, aggregations and sizes of SVs. Indeed, for cancers evolving in subpopulations of cells, where specific mutations may confer selective advantage for clonal expansion, more than half of the investigated cases harboured at least one clonal SV, thus suggesting that structural alterations could be highly implicated in tumour progression (Greaves 2015; Turajlic et al. 2019; Y. Li et al. 2020; Dentro et al. 2021). As highlighted below, both simple SVs or combinations of SVs within a locus, giving rise to a complex SV[8], concur in influencing cancer genome cell structure and tumour progression. In addition, subclusters of cancer cells may attain higher advantage and subsequently progress, with few mutational episodes

[8] Different types of complex events can be recognized by the characteristic signatures of formation embedded in them, they are: chromothripsis, extrachromosomal circular DNA (ecDNA), chromoplexy, breakage-fusion-bridge (BFB) cycles, aneuploidy and whole genome duplication (WGD).
Chromothripsis, as presented previously in this thesis, involves a single (or a few) chromosome shattering where many hundreds of rearrangements and losses occur after the fragmentation (Jan O Korbel and Campbell 2013; Hamdan and Ewing 2022). The ecDNAs are portion of shattered DNA that form circular replicated structures outside regular chromosomes, usually bearing oncogenes that can later be reincorporated into the chromosome (Storlazzi et al. 2010; Verhaak et al. 2019; Wu et al. 2019). Chromoplexy is defined as the joining of translocated and deleted portions of different chromosomes together (Shen 2013b). BFB cycles involve the formation of a dicentric chromosome resulting from loss of telomeres, fusing two chromosomes together (Gisselsson et al. 2000). Finally, aneuploidy results from lower or higher number of chromosomes than usual, while WGDs completely duplicate the entire genome of a cell (Holland and Cleveland 2009; Bielski et al. 2018).

represented by complex SVs, usually arising from single events, rather than by a continuous accumulation of mutations (Stephens et al. 2011; Jan O. Korbel and Campbell 2013; Shen 2013a; Anderson et al. 2018; Vendramin et al. 2021; Hamdan and Ewing 2022). The mechanisms by which SVs exert their pathogenicity usually involve gene dosage and gene fusion, commonly generated by simple (deletions and duplications) and complex (chromothripsis, ecDNAs, see footnote 8) SVs respectively, altering either oncogenes or tumour suppressor genes (Hamdan and Ewing 2022). Indeed, gene fusion rearranges closely together usually distant genes, as in the case of prevalent *BRC-ABLI* fusion in myeloid leukaemia (Druker et al. 2001; Hochhaus et al. 2017; Anderson et al. 2018; Cortés-Ciriano et al. 2020; Gerstung et al. 2020). Gene fusion is one of the interesting examples of diagnostic use of SVs, whose occurrence can be used as a proxy for the putative presence of certain cancers. Indeed, apart from myeloid leukaemia, structural alterations may signal the existence of prostate adenocarcinoma, Ewing sarcoma, glioblastoma or oligodendroglioma (Carver et al. 2009; Yao et al. 2014; Gorthi et al. 2018; Hochhaus et al. 2020a; Gonzalez Castro and Wesseling 2021). Another aspect of SVs that has been considered when addressing cancers is that the products of both gene amplifications and gene fusions are used as target for therapies treating different tumoral forms. Indeed, cases of haematological, lung, breast, ovarian cancer, T-cell lymphomas and neuroblastoma, among others, have undergone applications targeting either fused or amplified genes due to SVs as direct therapy (Sasaki et al. 2010; Hochhaus et al. 2020b; Q.-H. Li et al. 2020; Ewing et al. 2021; Kaushik Tiwari et al. 2022). Additionally, structural alterations can underlie specific cancer types but originate years or decades before the onset of symptoms, thus providing an extremely important role for early detection. Extreme cases concern the loss of chromosome 3p in clear cell renal cell carcinoma (predicted to arise 30-50 years before diagnosis) as well as in lung adenocarcinoma, where involved SVs emerge

decades before cancer onset (Mitchell et al. 2018; Lee et al. 2019; Oben et al. 2021). Overall, even if challenges still afflict research methodologies about detection and analysis of SVs in general, evidence has accumulated over their undisputable involvement in different aspects of cancer genomes, from early predictors of later oncogenic activity to targeting their products in therapies and evaluating their clinical relevance in tumour progression.

Neurodevelopmental disorders

Most of the research on CNVs and more broadly on general structural variations started focussing on pathogenic (often large) events significantly found in disease cases as compared to a set of healthy controls. Early reports, using different experimental techniques, started indicating that CNVs were found consistently in neuropsychiatric disorders such as autism spectrum disorder (ASD), schizophrenia and bipolar disorder (Harvard et al., 2005; M.-L. Jacquemont et al., 2006; Sebat et al., 2009; Sebat, Lakshmi, Malhotra, Troge, Lese-Martin, Walsh, Yamrom, Yoon, Krasnitz, Kendall, Leotta, Pai, Zhang, Lee, Hicks, Spence, Lee, Puura, Lehtimäki, et al., 2007; The International Schizophrenia Consortium, 2008; D. Zhang et al., 2009). These data, compelling despite being preliminary, indicated that SVs had a prominent role in neurodevelopment disorders (NDD). Classically, the broad category of neurodevelopmental disorders is defined as a set of conditions arising during development, such as attention-deficit/hyperactivity disorder (ADHD), ASD, learning disorders, intellectual disability (ID) and communication disorders with an association to a medical/genetic or environmental factor. More recently, studies addressing psychiatric disorders highlighted shared sets of risk alleles with ASD, ADHD and ID, suggesting the presence of a neurodevelopmental continuum among these syndromes emerging from altered brain development (Owen et al. 2011; American Psychiatric Association 2013; Owen and O'Donovan 2017; Morris-Rosendahl and Crocq 2020). CNVs

formed part of the risk alleles involved and the interplay among disorders emerged as these variants confer risk for different types of conditions, such as ID and schizophrenia, or that the load of CNVs in a genome correlates with the severity of the phenotype (higher in ID compared to ASD) and, additionally, highlighting the differential burden of large and rare CNVs among different NDDs (Girirajan et al. 2011; Kirov et al. 2014; Singh et al. 2017; Morris-Rosendahl and Crocq 2020). Given the involvement of CNVs in conferring a risk for different types of NDDs, it is relevant to consider this kind of genetic marker for early diagnosis and for advancing the understanding of the biological causes underlying morbid conditions. By using a large sample of trios, it was recently estimated that around 1 in 200 new-borns carry an NDD related CNV, providing key guidance for diagnosis and medical recommendations (Smajlagić et al. 2021). Moreover, not only copy number data uncovered previously undisclosed information about its influence over neurological development, but it also adds up to SNP data as a quantifiable source for improved diagnostic rate in trio pedigrees with NDDs, confirming once again the implications SVs play in human medicine and health (Zhai et al. 2021). Schizophrenia is defined as a complex psychiatric disorder typically showing hallucinations, impaired cognitive functions and amotivation symptoms; affecting less than 1% of individuals worldwide, it is one of the causes of premature death due to its harmful alterations (Whiteford et al. 2013; Owen et al. 2016; Charlson et al. 2018). Despite being a highly heritable disorder, with estimates of heritability around 80% in twins, this disorder is still diagnosed only via trait evaluation of psychiatric symptoms and no solid biomarker test exists to date (Sullivan et al. 2003; American Psychiatric Association 2013; Hilker et al. 2018; Kato et al. 2022). Research addressing schizophrenia using CNVs uncovered a previously unknown contribution of SVs over the risk for such disorder, with both common and rare alleles contributing to the associations (Rees et al. 2014; Tansey et al. 2016; Marshall et al. 2017)

and specific deletions (22q11.2 and 3q29) being highly associated as risk factors (>50 fold) for this disorder (Stankiewicz and Lupski 2010; Kato et al. 2022). Intriguingly, the 22q11.2 deletion is seemingly associated with lower expression of protein kinase R-like endoplasmic reticulum kinase (PERK), leading to decreased protein synthesis, endoplasmic reticulum stress and abnormalities in F-actin functions; these phenotypes were partially restored in neuron cells by the active pharmacological control of PERK action (Arioka et al. 2021). Similarly, a deletion in *ARHGAP10* gene putatively activates the RhoA/Rho-kinase signalling pathway[9] (Kato et al. 2022) and a study indicated that neurons harbouring this deletion showed alterations in branching number and neurite length. Nonetheless, once again the physiological state of these features was restored by the addition of Rho-kinase inhibitor (Sekiguchi et al. 2020). As mentioned above, a set of pathogenic CNVs can have comorbidities among different NDDs and individuals having such variants show higher resistance to medical treatment against schizophrenia symptoms, highlighting that the severeness of psychiatric outcomes are more serious in subjects with higher number of clinically relevant CNVs (Kushima et al. 2017; Sobue et al. 2018; Kato et al. 2022). On the other hand, medical treatment targeting pathways disrupted by pathogenic variants can help recover lost functions and restore normal physiological activities, providing important insights for the CNV mechanisms of action in schizophrenia and improved therapy outcomes. ASD comprises a heterogeneous set of neuropsychiatric conditions involving limited and repetitive behaviours, interests and activities, impaired social communication and interaction, often accompanied with ADHD, anxiety, depression and epilepsy (Lord et al. 2020). It is a highly

[9] The RhoA/Rho-kinase pathway is involved in neurite outgrowth, providing neural migration dendrite development and axon extension, thus playing a role in the pathophysiology of central nervous system diseases (Fujita and Yamashita 2014; Xiang et al. 2021)

heritable disorder as shown in twin studies, with its prevalence increasing in different countries and, as a USA report indicates, ASD affects 1 in every 68 children with higher estimates in boys than girls (Hallmayer 2011; Ronald and Hoekstra 2011; Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators and Centers for Disease Control and Prevention (CDC) 2014; Takumi and Tamada 2018). Different studies reported significant associations of CNVs in ASD subjects, also making use of large family datasets, and helped characterize in finer detail the contribution that SVs have upon the disorder, identifying many loci of interest (Pinto et al. 2010; Griswold et al. 2012; Sanders et al. 2015; Leppa et al. 2016). Using large datasets researchers narrowed down the most frequent copy number alterations identified in affected individuals: 2p16.3, 16p11.2 deletions and 15q11-13 duplication; with up to ~7% of the affected subjects carrying CNVs including chr16 deletion and chr15 duplication (Pinto et al. 2014; C Yuen et al. 2017). Animal models harbouring 16p11.2 deletion indeed resemble human ASD traits, with motor/memory deficits, repetitive behaviours and hyperactivity. In fact, a number of genes at this locus, and their related pathways, are involved in neuroanatomical phenotypes regulating synaptic transmission, dendritic formation and arborization, implicating this locus as one important contributor to ASD phenotypes in humans (Horev et al. 2011; Calderon de Anda et al. 2012; Golzio et al. 2012; Portmann et al. 2014; Arbogast et al. 2016; Blizinsky et al. 2016). CNVs encompassing 15q11-13 overlap numerous genes with intriguing characteristics; chromosome 15 is indeed prone to copy number alterations and the long arm region q11-13 harbours five common breakpoints for CNVs, the first of which hosts deletions involving brain functions (Bailey et al. 2002; Stefansson et al. 2014; Takumi and Tamada 2018). This first region produces the cytoplasmatic FMR1-interacting protein (CYFIP1), responsible for cytoskeleton regulation, translation at synapse and may be involved in neural structures

development (Napoli et al. 2008; De Rubeis et al. 2013). On the other hand, studies revealed that duplication at this locus show enhanced dendritic density and spines, showing that plasticity in copy number at this region reflects directly on the underlying brain functions (Oguro-Ando et al. 2015). Another case of plasticity can be detected in the second region, where deletions are usually responsible for patrilinear and matrilinear-inherited disorders, namely Prader-Willi and Angelman syndromes respectively, but where duplications have also been linked to ASD (Depienne et al. 2009; Takumi 2011). A noteworthy characteristic of 15q duplications is that children with both this CNV and ASD exhibited higher impairment in motor and daily living skills, compared to children lacking the variant and syndrome, suggesting the putative biomarker use of motor impairment for the duplication (DiStefano et al. 2016). The last region contains two major deletions associated to NDDs, specifically to ASD, developmental delay, epilepsy and schizophrenia; these allegedly causative deletions categorize in the medium to large size range being ~1.5 Mb and ~680 Kb respectively (Sharp et al. 2008; Shinawi et al. 2009; Ionita-Laza et al. 2014). Taken together, the knowledge gathered so far about CNVs and NDDs shows how vast and interconnected the variant-disorder system really is, with the extent and function of the genomic elements encompassed by deletions and duplications depending on their magnitude. Undoubtedly, specific key genes whose disruption is directly connected to impaired brain development exist, however the intricate network of regulatory elements, gene products interactions and how disruptive a CNV can be, all play a role in determining the actual burden of this variants upon NDDs.

Summarizing, copy number variation has revealed its substantial contribution to a number of conditions affecting different aspects of health in humans. Their action can directly alter the normal function of genes involved in drug metabolism or alter physiological process, concurring to a

number of implications, as exemplified for NDDs. They also seem to constitute a “physiological” property of the genomes of cancer cells, but can nonetheless be used as a diagnostic marker, whose occurrence may suggest the presence of certain types of cancers. Overall, CNVs play an important role for different aspects of human health and further understanding of the mechanisms they participate can also help refining early screening in clinical diagnosis.

1.2 Addressing human population genetics using structural variation

1.2.1 Human evolution through the lens of population genetics

Research in evolutionary genetics sets its foundations on the fact that, when studying an extant organism, part of the past events that shaped its genetic configuration occurred in the genomes of its ancestors. In a narrower context, human evolutionary genetics aims at comparing current variation among human beings to highlight differences that might reveal the causative actions of past dynamics, which were responsible for such variation. These dynamics can be the result of simple mutational events, of different selective pressures (and thus responses) that our ancestors had to cope with, or reflect how past populations grew, shrank, split, or merged with each other to pass down these signatures to future generations. At a population level, natural selection will thus favour individuals carrying those variants that allow a better adaptation to the environment they inhabit and consequently have higher chances of transmitting them. The interchange of individuals among populations, defined as gene flow, will also have an impact on the relative frequencies of variants within populations. Also, the number of individuals in a population and its change in size will affect the amount of variation through a process whose contribution was acknowledged only a few decades ago, genetic drift. Genetic drift can be summarized as the random changes in the distribution of an allele in a population, mainly due to the chance by which the members of the population pass their alleles to the subsequent generation. This process, through time, ultimately leads to the fixation or loss (100% vs 0% frequency respectively) of an allele in the population. The empirical data showing a much higher diversity of polymorphisms in humans than expected led to the development of the neutral theory of molecular

evolution, where the appearance or maintenance of mutations is largely due to genetic drift (Kimura and others 1968; Nei et al. 2010). Ultimately, the major forces capable of influencing the frequency of an allele in a given population are those summarized above, selection, gene flow, mutation and genetic drift (Relethford and Harding 2001). Additionally, the human species is widespread on Earth, intrinsically inhabiting different environments and highly variable in its cultures. Thus, the evolution of humans can be viewed with greater clarity when considering the concerted action of all these dynamics together (Creanza and Feldman 2016).

Since the marked spread of high throughput sequencing techniques, sparked by the promising source of information due to the abundance of SNPs in a human genome, the majority of the research focussed on variation at single nucleotide to assess human diversity and evolution, significantly favoured by highly specialized platforms for massive detection (Collins et al. 1998; Kruglyak and Nickerson 2001; LaFramboise 2009). The wide interest in SNPs is also reflected by the sharp increase in deposited variants within the context of public datasets, which, in the early 2000s, increased five-fold (Smigielski et al. 2000; Nielsen 2004). In general, this interest flowed in different specific lines of research, focussing on the first reconstruction of co-occurrence among SNPs (haplotypes), inferring ancestry, demography and admixture in populations (Reich et al. 2001; Wakeley et al. 2001; Clark et al. 2003; Hellenthal et al. 2014). These emerging discoveries depicted a progressively clearer picture of past events and relationships among humans, but also how such dynamics could leave quantifiable information in the genome. It was not long before researchers developed and deployed methods to infer the presence of selective events at specific loci and how their presence nowadays can impact health (Sunyaev et al. 2000; Sunyaev et al. 2001; Sabeti et al. 2002).

1.2.2 Something new on the horizon: structural variants in population genetics

As reported in the chapter on structural variants, growing interest in the scientific community for this type of mutations fostered progressively complex research of SVs in humans. The SV information that could be initially gathered from humans was mainly restricted to macroscopic events in single individuals and lacked a proper population-scale characterization. A pioneer study used wet lab approaches to detect large SVs in a small group of individuals, mainly Europeans but also including Chinese, Native American, Indo-Pakistan and Sub-Saharan Africa samples, provided encouraging results. The authors found an average of 12 CNVs in each sample and more than the half of all the variants found overlapped with coding regions, triggering interest on their possible role in disease formation or control over gene expression (Iafrate et al. 2004b). These interesting new findings kept the attention of researchers alive and triggered ever-complex investigations, aiming at filling the gap concerning the lack of population-scale information about SVs. The first innovative study involving a larger dataset, specifically including different populations, addressed almost 300 samples from the HapMap dataset having European, Yoruba (African), Japanese and Chinese ancestries. Upon integration of two different platforms, SNP array and aCGH, (Redon et al. 2006) found an average of ~100 CNVs per genome that were merged in some 1400 copy number variable regions (CNVRs, genome regions hosting merged CNVs) covering more than 300Mb of the human genome and mainly composed of deletions, duplications, multi-allelic events and mixed deletions/duplications. The authors provided particularly interesting information, such as that deletions intersecting genes seem to be negatively selected; they highlighted which functional categories of genes are enriched in CNVs and proposed that these play a role in disease due to their

intersection with known pathogenic genes. A peculiar achievement from this study was also to find population clustering using CNV data, thus showing how samples from different populations could be discerned using these markers. In particular, they devised the V_{ST}[10] metric showing population stratification by CNVs, matching previously known differentiated loci. Studies specifically focussing on the putative implication of this newly investigated set of variants hypothesized that the gene categories with signals of selections that were enriched in SVs might also be explained with a relaxation of constraints at these loci as attested by the accumulation of copies in olfactory receptor genes (Nguyen et al. 2008; Young et al. 2008). Further assessment of different continental samples from the HapMap project found interesting differences of frequencies for a deletion encompassing the *UGT2B17* gene, implicated also in the metabolism of steroid hormones (Xue et al. 2008). In particular, it was highlighted how different human populations carried different proportions of the deleted allele and exhibited specific patterns of selection in the affected genomic locus; signs of positive selection were found in East Asia, while the maintenance of both alleles within Europe matched signatures of balancing selection. Thus, an exciting new layer of information started accumulating regarding the distribution, evolutionary significance and possible impact on health of SVs among worldwide human populations.

Itsara et al. (2008) relied on SNP arrays to characterize CNVs in a large scale dataset containing around 2500 samples and described how large (> 500Kb) variants were indeed quite common in worldwide populations, reaching up to a frequency of 10%. This study also highlighted how,

[10] The authors conceived this metric to assess the variance in copy number between pairs of populations. They analysed the $\log_2$ intensity rations from the platforms used (aCGH and SNP array) and computed $V_{ST} = (V_T - V_S)/V_T$, where V_T represents the variance in $\log_2$ ratios among all samples from the two populations, and V_S is the average variance in $\log_2$ ratios in each population, scaled for its size.

although present in the population, large variants are generally depleted of gene-rich regions, found variants intersecting genes that contribute to known disorders and emphasized how a number of events may signal putative genetic diseases of interest. Another important approach was the inclusion of different worldwide populations leveraging whole genome sequencing data, which was scarce at that point. Mills et al. (2011) combined different populations and short-read sequencing to infer and analyse CNVs in a large dataset, addressed the samples from the pilot phase of the 1000 Genomes Project (1KGP) and initially identified thousands of deletions and duplications. This was one of the first occasions in which continental groups were studied with short-read sequences and it was not until the complete sequencing of more than 2000 genomes that a worldwide view of SV variation was consolidated. Known as the Phase 3 of the 1KGP, complete sequence information of ~2500 samples from 26 populations was available, this research, in the study by Sudmant, et al. (2015), identified more than 2000 SVs per genome and estimated that their influence over gene expression was up to 50-fold greater compared to SNPs. The study had sufficient power to confirm previous data suggesting stratification of SVs among populations and highlighted how selected variants might have had an adaptive role in the adaptation to novel environments.

Although it is likely that genetic drift had probably played a major role in shaping the apportionment of SVs present in humans nowadays, intriguing examples highlighting the possible action of positive selection exist. The case of α-globin previously reported is an interesting illustration: most humans indeed carry two copies of *HBA1* and *HBA2* genes, but deletions at this locus seem to be maintained at higher frequencies in the sub-Saharan regions of Africa, due to their putative protective actions against malaria (Kan et al. 1975; Williams et al. 2005; Lam and Jeffreys 2007). Furthermore, recently attention has been focussed on another locus where

duplications of glycophorin genes *GYPA*, *GYPB* and *GYPE*, mainly found in the oriental part of Africa, may have emerged due to their protective action against malaria and may have spread recently because of positive selection (Leffler et al. 2017; Louzada et al. 2020). Lastly, examples of adaptation likely driven by SVs involve the *HERC2* gene locus, associated with skin colour, where a duplication seem to be under negative selection in Europeans, or an inversion mostly found in European samples whose genomic region (encompassing the *KANSL1* gene) shows signs of positive selection and seems to be implicated in women's fertility (Gudbjartsson et al. 2005; Boettger et al. 2012; Saitou and Gokcumen 2019b; Almarri et al. 2020).

The appreciation for the role of SVs in human health prompted three initiatives to create repositories for their variation: i) the DECIPHER project, launched as an open online repository of genomic variants and their related phenotype with the intent of advancing the understanding of the clinical relevance of CNVs (Swaminathan et al. 2012); ii) the Deciphering Developmental Disorders study (DDD), launched in 2011 were data about 12000 children with undescribed developmental disorders was deposited to further characterize them and their parents and improve diagnosis using different genome analysis platforms, collecting all types of genetic variation including SVs (Firth and Wright 2011); and iii) the Database of Genomic Variants (DGV), launched in 2006, has been an online resource for cataloguing SVs in the human genome and, around the inception of DDD and DECHIPHER projects, DGV already contained more than 55 published studies summing up to 22000 genomes and around 2.5 million entries (Zhang et al. 2006; MacDonald et al. 2014). These repositories served as integrated collections of all SVs reported in literature to generate a comprehensive open-source database for both global variation in humans and pathogenic variants. Such platforms aided other researchers in the

identification of newly discovered SVs or confirming already deposited ones, thus enabling an ever-growing map of genomic rearrangements in our species. On the clinical side of these databases, they assisted a more rapid identification of the implicated phenotypes for an identified variant, as previously undescribed disease-causing alterations deposited there allowed to widen the horizon of clinically relevant alterations and thus accelerating the assessment processes in medical genetics studies.

In another important study for the analysis of the population genetics of structural variation in humans, over 200 individuals from 125 populations worldwide were investigated, and it was uncovered how deletions seem to be under stronger selective pressure than duplications (Sudmant, Mallick, et al. 2015). In fact, deletions better recapitulate population structure, because of the tendency of duplications to mutate faster, thus homogenising variability among individuals and populations, concealing previous events and making similar variants more prone to be identical by state rather than by descent (Sudmant, Mallick, et al. 2015). In this project, also ancient genomes from Neanderthal and Denisova were analysed, and in agreement with other studies on SNP genotypes, a duplication was found to be private (reaching high frequencies) in present day Oceanian individuals and only shared with the Denisova hominin. About this last point, subsequent investigations found further evidence of SV-specific signatures of introgression/adaptation between archaic and extant hominins. One example involves a human-specific expansion and differentiation, with respect to archaic hominins, for a locus potentially involved in adaptive responses to dietary or environmental temperature shifts (Hsieh et al. 2021). Subsequently, ~20 CNVs were found to show signals of positive selection likely resulting from the introgression of archaic hominins into the ancestors of current Melanesians and, specifically, two complex rearrangements in chr16 and chr8, absent in other human populations

(Hsieh et al. 2019). At 16p11.2, the authors resolved a large duplication introgressed from Denisova to the ancestors of present-day Melanesians around 170-60 kya and present at high (~80%) frequency. This duplication bears signs of positive selection and resides close to a locus where structural rearrangements predispose to autism disorder. The SV at 8p21.3 is instead more complex, being composed of a 38kb duplication and a 6kb deletion, introgressed in Melanesian from Neanderthals around 120-40 kya, showing signs of a selective sweep and thus indicating that selective forces most likely shaped its apportionment and preservation in the population.

When long-read technologies were released and started to be used in genomic research, the first results of the attempts to leverage their strengths to resolve in finer detail SVs in humans started to emerge. As discussed previously, long-read sequencing greatly improves the detection of large SVs, novel insertions and resolution within repetitive segments of the genome, thus, with these promises, it is not surprising that many studies ventured into the analysis of structural rearrangements using this new technology of genome analysis. In one of the first studies that incorporated PacBio sequencing to refine the discovery of SVs in the human genome, Chaisson et al. (2014) detected ~26000 SVs resolved at base-pair level containing difficult events for short-read technology such as long stretches of tandem repeats and complex insertions. These initial results underlined the ability for PacBio to reach precision peaks for variants around 5kb and be particularly well-tuned for previously difficult SVs in the form of complex insertions and repetitive DNA sequences. Innovative approaches also integrated different sequencing technologies, like aCGH, short-reads, long-reads among others, to obtain almost 10,000 SVs between 100bp and 1Mb in length and spanning ~60Mb of the reference genome, of which 1/3 were exclusively identified with PacBio; hence showing the power offered by using multiple sequencing technologies and specifically by long reads

(English et al. 2015). The first analysis of Asian samples with long reads also confirmed that this methodology can be a source of new genomic information not only confined to European ancestry. The sequencing of an individual of Chinese ancestry revealed ~10,000 SVs (deletions and insertions) absent from the reference and mostly or mobile element insertions (MEI); further assessment of a Korean individual, via multiple sequencing technologies, also highlighted ~18,000 unreported SVs among which many insertions were shared across Asian populations (Seo et al. 2016; Shi et al. 2016). These new results not only confirmed the known strengths and new opportunities offered by long-read approaches, but also shed light on hidden SV variability among populations which, so far, were still uncovered. Two haploid human genomes were sequenced with PacBio, leading to the discovery of ~26000 novel SVs that were not present in the Phase 3 1KGP and, as expected, most variants were < 1kb in size (Huddleston et al. 2017b). Intriguingly, the authors showed that, upon performing short-read sequencing on the samples analysed and using SV calling algorithms for this type of data, 61% of the SVs found with PacBio could be accurately genotyped, meaning that separating calling (PacBio) from genotyping (Illumina short-reads) still allows to correctly assign a genotype to most of the variants not found by Illumina sequencing. This fact may facilitate large population-scale studies where budget is a limiting factor, since this research might rely on a few samples from different populations for the discovery stage using PacBio (thus limiting the costs), while using Illumina sequencing on hundreds or thousands of samples to complete the set and still recover accurate genotypes for subsequent analyses. Further progress in the detection of SVs using PacBio was made by a long-read sequencing project analysing 15 deep-coverage genomes (~57X) and found almost 100,000 variants (insertions, deletions and inversions), most of which laid undetected in the 1KGP or other databases (Audano et al. 2019). Around 2,000 SVs shared among all samples were

not present in the current reference genome, highlighting how the current human reference is not representative of the global variation and still may contain errors or minor alleles at SV loci (Audano et al. 2019; Ho et al. 2019). Remarkably, the authors showed how almost half of the detected variants intersect either genes or regulatory elements and how a surprising amount (55%) of variable number tandem repeats (VNTRs) preferentially map to the terminal 5Mb of chromosomes, representing a nine-fold bias toward these locations.

More recently, samples from worldwide populations including ~18,000 individuals were short-read sequenced (also validated with long-read technologies), and more than 200,000 SVs were revealed in the callset with an average of 4,000 per genome, identifying more than 20,000 gene-altering SVs, mostly comprised of deletions, of which ~10% alter three or more genes and might be hundreds of times more disruptive than rare SNVs (Ho et al. 2019; Abel et al. 2020). This study additionally showed that over 300 individuals carry ultra-rare and Mb-size SVs which may have large impacts on their health, thus guiding further analyses on the functional interpretation of worldwide SV variation in human populations. Other vast sources of information over mutations altering protein coding genes began accumulating with the Exome Aggregation Consortium (ExAC), where 60,000 exomes served as a catalogue to identify pathogenic variants and detect genes depleted of protein-truncating mutations (Lek et al. 2016). The project helped in defining such sets of genes showing a strong selection against damaging mutations, and thus putative new disease phenotypes were inferred. A subsequent expansion of the ExAC came from the Genome Aggregation Database (gnomAD), where around 125,000 exomes and over 15,000 genomes from human sequencing studies were gathered to deeply analyse gene disruptive variants and classify them for their tolerance to loss-of-function, thus improving characterization for common and rare

diseases (Karczewski et al. 2020). A branch gnomAD project specifically addressed SVs for medical purposes and worked on a set containing over 12,000 samples from global populations (54% of non-European ancestry) from which they retrieved ~433,000 SVs including deletions, duplications, insertions, inversions, translocations, multi-allelic and complex CNVs (Collins et al. 2020). The authors, analysing this cohort composed of 46.1% European, 34.9% African/African American, 9.2% East Asian, 8.7% "Latino" and 1.2% of admixed/other populations samples, retrieved an average of ~8,000 SVs per individual (more than twice compared to former projects), remarking the advantages of joint high-coverage data and robust calling pipeline when investigating worldwide populations coming from different studies. Moreover, one quarter to one third of all protein-disrupting mutations were caused by the SVs identified in this study, highlighting an exceptional worldwide contribution for variants altering the function of genes and the need for medical genetics and pharmacogenomics to address this type of variation when investigating disease-related phenotypes and designing the action of drugs.

The study of structural variation within specific isolated population still represents a minor branch of the research in this field. Still, as highlighted so far, the investigation of an ever-growing number of individuals from worldwide distributed populations cyclically recovers previously unknown variants (usually rare ones), sometimes specific to certain groups. Throughout human history, we acknowledged the existence of previously unseen microscopic organisms in freshwater, proved the existence of the finest inseparable grains of matter as well as the characteristics of our solar system and galaxy, only to find out that greater levels of complexities exist for us to be discovered when using more and more sophisticated methods of analysis. This inherent and universal characteristic of research also applies to the complex network of variation in SVs among human

populations which, despite being insufficiently explored, produced some interesting results when underrepresented or isolated population were studied. We need to specify that underrepresented populations, as the term itself expresses, are all those groups that are usually understudied or not studied at all. Due to the history of research in the scientific community, this term equals to populations of non-European descent. Research, being historically conducted in Western (over)developed countries, namely Europe and United States of America, made it a consequential and easier choice to start gathering information from quickly available samples. Also, studies whose final output only concerns the world where the authors of such research live in gathered "relevant" information, for example in genome wide association studies, for disease or traits mainly (if not only) present in this leading population of European ancestry, while disregarding humans living in different geographical locations (Bustamante et al. 2011; Popejoy and Fullerton 2016). This approach led us to this unconventional and biased representation of human genetic and phenotypic features, where our advancements in picturing an increasingly clearer idea of global variation and gathering knowledge on health-related variants were often placed in a Eurocentric perspective (Need and Goldstein 2009; Zeggini 2014; Sirugo et al. 2019).

An early study focussing specifically on Chinese individuals concentrated on CNVs affecting height in a set of ~600 samples and although the corrected results did not meet statistical significance, the authors found gains and losses indicative of possible trends in height variation (Li et al. 2010). Intriguingly, the two CNVs most related to the measured height among subjects, reside in genome loci previously reported to influence height or have height-altering features. Research on three European isolates, thus maintaining the focus on the most represented ancestry but peeking into specific isolated communities, highlighted both general CNVs

frequency sharing among populations as well as within-isolate relatedness for specific variants, emphasizing how the population-specific dynamics occurring among isolates leave a detectable trace and differentiate populations (W. Chen et al., 2011). Another investigation in the Finnish population showed specific apportionment of CNVs in this population with respect to African or European ancestry individuals and also novel variants that further illustrate the landscape of structural variation in this population isolate (Kanduri et al. 2013). CNVs were also used as proper ancestry predictors for three minority ethnic groups in China. Indeed, admixture patterns within the dataset (on its two geographical extremes composed of Europeans and Chinese samples) specifically devised to analyse Uyghur, Kazakh and Kirgiz populations recovered intermediate to high proportions of Chinese and European components respectively, confirming their intimate history of relationships between Eastern and Western Eurasia in the context of historical connections dating back thousands of years ago (Lou et al. 2015). As stressed before, isolated or underrepresented populations are a source of information often disregarded. In this study, the authors report population-specific CNVs that might also have functional effects, with many variants particularly enriched in genes implicated in wound response. Other works on specific populations highlighted how 250 Dutch families as well as ~1,000 Japanese samples had population-specific SVs and SVs hotspots previously missing from larger datasets like 1KGP (Kloosterman et al. 2015; Nagasaki et al. 2015; Hehir-Kwa et al. 2016b; Ho et al. 2019). More recently, attention has risen for underrepresented and isolated populations and some interesting reports addressed specific African communities and Malaysian indigenous populations. These investigations found signals of selection for copy number variable loci, specifically for these populations, at genes responsible for putative adaptive phenotypes such as drugs-toxins metabolism and immune response (Fu et al. 2018; Nyangiri et al. 2020), digging out evolutionary relevant variants

that diverged in frequencies and consequently highlighting the differential selective pressures and adaptive responses, among humans, in an SV framework.

More recently, interest in pharmacogenetics[11] paved the way and sparked interest for the development of different studies on CNVs, particularly for their involvement in different disorders, as discussed before. A team effort gathered evidence to assemble a pharmacogenetic genotyping panel and provided over 100 variants including CNVs with interindividual drug response variability for comprehensive multi-population screening (Scott et al. 2021). Conversely, other authors focussed specifically on underrepresented groups and delivered the first representation of CNVs variation in genes involved in pharmaceutical treatment within the Colombian population (Ramírez et al. 2019). The research highlights greater CNV variability in the *GST* gene family than in cytochrome P-450, respectively implicated in the detoxification of carcinogens agents, therapeutic chemicals, environmental toxins for the former, and in the phase I metabolism of most medications for the latter. Even though the authors state that still more progress has to be made to uncover the functional consequences of the identified CNVs, this first Colombian pharmacogenetic investigation still represents a valuable and novel source of information for representing the variability of response to drugs in humans.

[11] Pharmacogenetics is a multidisciplinary term describing the study of the variation of drug responses that are attributable to genetics, thus traceable to features that are inherited, that can vary among individuals (or populations) and can be responsible for differential outcomes to the same treatment.

Even if historically disregarded, research in human structural variation has increasingly revealed the different roles and aspects these variants play in many contexts. From implications in adaptive processes during human history to the action on medically relevant traits, structural rearrangements revealed an interesting layer of information previously underestimated, whose implications in the understanding of our evolutionary history and health open new perspectives from which we can address future research. On a broader scale, variation among both individuals and populations, and the ever-increasing refinement of detection methods uncovered a previously undescribed mutational landscape of genetic features that must become a routine in human-based genetic screening if researchers want to further address the complexity of genomes. A substantial part of the research leans on the clinical aspects of SVs, due to their involvement in diseases and also the preventive diagnostic uses offered by known associations or mechanisms of action. Finally, also their action on loci of pharmacological importance need to be further elucidated to keep including the action of SVs, notoriously spanning larger sequence stretches than point mutations, in these context and offer proper treatments to patients differing in their genomic structural architecture.

1.3 The Romani people

1.3.1 "Oh, Romani… the citizens of Roma in Italy, right?"

The Romani (or Roma[12]) population is generally defined as the ensemble of groups historically originated from India and afterwards dispersed throughout all Europe, sharing a common heritage of culture and traditions. Even though, since historical times as well as nowadays, Romani presence is not restricted to Europe, it is precisely in Europe that they are often referred to as the largest transnational minority ethnic group (Commissioner for Human Rights 2012). This apparently contradictory term, being a numerous yet minority group, is easily explained by the fact that 12 million Romani people are estimated to reside in Europe, however their presence across countries largely fluctuates and never reaches 10% of the population of any country (Commissioner for Human Rights 2012). The large geographic variability is also mirrored by the diversity in terms of language: within the Romani population, a set of more than 60 dialects exists, generally clustered into four groups: Central, Northern, Vlax and Balkan; belonging to the New Indo-Aryan branch from the Indo-European languages group (Bakker 2000; Hancock 2002; Matras 2002). On its own, the development and diversification of Romani language reflects the complex population history of this group, where encounters and cultural interchange within different countries had an influence in shaping the current variation. Moreover, some dialects are the product of the local language spoken in a country plus the addition of Romani words, making up a para-Romani dialect (Fraser 1992; Hancock 2002). Examples of these para-Romani dialects are the Angloromani in the United Kingdom or the

[12] Romani and Roma are equally accepted ethnonyms for referring to this population and may be used interchangeably but, for the sake of clarity, some groups do not identify as Roma because in some dialects specifically refers to "married Romani man" and not a general member of the Romani population (Hancock 2002).

Caló in Spain (Hancock 1984; Bakker 1995; Matras et al. 2007). As stated above, most Romani groups reside in Europe; nevertheless, for historical reasons many individuals left (or were forced to leave) and, nowadays, other groups live in the Middle East, United States of America, Canada and Brazil (Fraser 1992; Kenrick 2007). Despite these types of variability, the Romani population is considered one single entity with its own identity and traditions, all sharing the same geographical origin in India and westward dispersal routes, making it a single ethnic group (Fraser 1992; Hancock 2002).

1.3.2 History so far

Before dwelling on the more biological aspects of the population genetics of Romani, which indisputably confirmed and revealed a great amount of insights on this group, it is relevant to put in context the history of this population, where they originated and what has been their subsequent diaspora, the groups and countries they interacted with and how those groups received and treated them along the centuries, up until one of the darkest pages of modern history with the advent of Nazi regime and the Holocaust.

As stated above, despite controversies, differing hypotheses and lack of a self-written records of Romani making, there is a consensus on an Indian origin of Romani. The first evidence came from linguistics as early as the 18th century, where studies identified a number of analogies with Indian languages (Fraser 1992; Hancock 2002), further supported by specific similarities with both central and north-western languages (Turner 1926; Matras 2002). Cultural anthropology studies also found similarities with some Indian social structures where work affiliation often differentiate different clans, that usually show endogamy practices (Fraser 1992; Iovita Radu P. and Schurr 2004). It has been postulated that the original proto-Roma population could have been related to nomad groups residing in India

that had inherent musical traditions, in agreement with later reports of musical practices in Romani groups in Europe once they left India (Pott 1844; Fraser 1992). Additionally, some proposed that ancestral Romani belonged to military clans in the north-western part of India (mainly Rajasthan) while others hypothesized a relation with nomad groups outside India that speak Indo-Aryan languages, like Romani do (Sampson 1923; Kochanowski 1968; Hancock 2002).

It is generally acknowledged that the proto-Roma population residing in north-western India, at some point in time migrated north-westwards through Persia, but the lack of self-recorded history still affects the certainty of the dates for this event, the exact major event that initiated their diaspora, in part responsible for the current apportionment of Romani groups in Europe. Despite the lack of an exact time frame for their exit from India, northern Indian regions experienced periods of tensions and wars with neighbouring empires and the presence of displaced groups has been historically documented (Fraser 1992; Hancock 2007). Different hypothetical dates for out of India event exist: one derives from a story stating that around the 5th century, the king of India Shangul donated over 10,000 musicians from India to the Persian king to amuse his army, matching the hypothesis of the proto-Romani being associated to groups with musical expertise (Fraser 1992; Kenrick 2004). Another possibility argues for the abandon of India around the 1000 CE due to the southeast expansion of the Ghaznavid empire; those periods of conflicts might have triggered groups to leave the area either to escape from or fight the war (Fraser 1992; Hancock 2002).

Once the proto-Roma left India, they probably quite rapidly reached Persia and current Armenia, thus entering the Byzantine empire (Fraser 1992; Marushiakova et al. 2001; Hancock 2002; Kenrick 2004) and even though dating such events remains a hard endeavour, once again the study of

language can help infer the routes used by proto-Roma to keep heading westwards. Indeed, the lack of many Arabic words and, conversely, the presence of more Greek ones suggests a more likely migratory route in Anatolia and a longer interchange (either in time or cultural influence) within the Byzantine Empire (Sampson 1926; Matras et al. 1997; Matras 2002). The European presence of the Romani probably started after the 13th century, when some groups reached the Balkan peninsula, others moved south to the Peloponnese[13] and later, during the 14th century, some other went north towards the Wallachia principality (Fraser 1992; Marushiakova et al. 2001; Hancock 2002; Kenrick 2007). Within these territories the first imposed slavery measures were taken to their regards, enforcing them to either serve people or perform hard works as farmers; those who were freed or escaped these legal liberty-restraining measures kept migrating further west, reaching central and western Europe (Hancock 2002; Kenrick 2007).

In these periods, we have the first written records from host countries describing the arrival of Romani groups or specific happenings concerning them often describing the arrival of groups of musicians, blacksmiths, skilled in horsemanship and farmers from an unspecified "Little Egypt" (see footnote 13) and, at first, they were accepted in many cities across Europe (Fraser 1992; Kenrick 2007). About their origin, during the 14th century confusion was created by the spread of this "Egyptian" affiliation from which other names subsequently derived, such as "*gypsies*", "*tsigani*", "*zingari*", "*zigeuner*" or "*gitanos*" (from England, Romania, Italy, Germany and Spain respectively), to the point where Europeans learned to recognize them better by these externally imposed names rather than by their self-given/recognized name Romani (as ironically exemplified by the

[13] An area in the Peloponnese, close to the town of Methóni, was allegedly known as "Little Egypt", which is also the toponym of the hometown given by the first Romani people to the encountered citizens when entering Western European countries (Fraser 1992).

title of the first paragraph in this chapter) (Fraser 1992; Hancock 2002). During the 15th and 16th centuries, Romani reached virtually all European countries, England and Russia included (Figure 10), but the first welcoming behaviour started to change; Romani groups were not allowed to cross specific lands, tailored laws were issued specifically for them and belonging to this ethnic group caused automatic expulsion, a general anti-Romani sentiment permeated Europe (Fraser 1992; Hancock 2002; Achim 2004; Kenrick 2007).

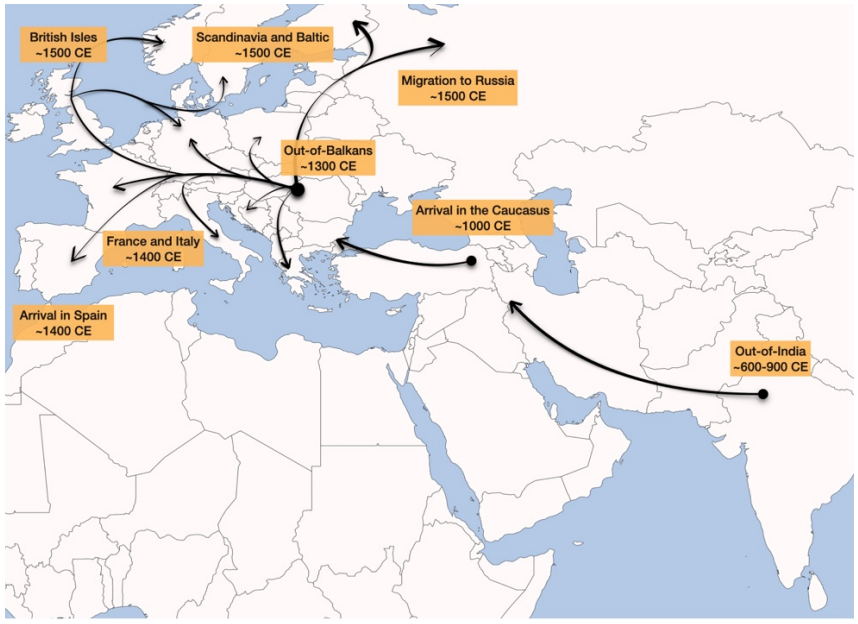


Figure 10. Romani diaspora. Simplified scheme of the Romani diaspora from Northern India to European countries. Arrows indicate main common directions of dispersal; banners highlight geographical regions and dates of arrival. Data form (Fraser 1992; Kenrick 2007; Mendizabal et al. 2012; Moorjani, Patterson, et al. 2013).

Moreover, the climate of radical change in politics and clashes within European countries probably did help this general intolerance that forced them to imprisonment, forced settlement or expulsion, enslavement, deportation[14] and death sentences (Boyd-Bowman 1985; Fraser 1992; Barany 2001; Brearley 2001; Hancock 2002). Periods of forced unification within the countries Romani were residing followed throughout the 17th/18th centuries, trying to blend diversities together in light of an imposed homogenization with the ruling and empowered realms. Moreover, most groups lived as slaves in many parts of Europe as late as 1855, when slavery ended (Marushiakova et al. 2001; Hancock 2002; Kenrick 2007). Unfortunately, their situation did not improve due to the newly abandoned slavery practices in different countries, as stigma and persecutions continued, culminated with the Romani Holocaust perpetrated by the Nazi Germany regime during the World War II, estimated to have erased the lives of hundreds of thousands[15] Romani people (Lutz 1995; Lewy 2000; Sridhar 2006). Additionally, after the end of the war, the Romani were not recognized as victims of a genocide, disregarded once again even in front of their collective mourning (Lewy 2000; Sridhar 2006) and about 40 years had to pass before the official recognition arrived.

Currently, Romani people are considered citizens of the European Union and during the 1970s different organizations were born to promote inclusion, emancipation and legal recognition of this ethnic group and in 1971, London hosted the first World Romani Congress, attended by representatives of nine nations to assess different issues (common culture, language and crimes, among others) and established the official Romani

[14] During this period, in 1538, the first deportations to the Americas took place in Portugal

[15] Apparently, there is no consensus or precise data on the total amount of Romani individuals killed in the genocide: estimates range from around 190,000 to 500,000 victims (Lewy 2000; Sridhar 2006).

flag (Figure 11). Nowadays, despite the international recognition and support, stereotypes and prejudices from non-Romani citizens still undermine an equal and peaceful coexistence and living, as Romani still lack proper levels of employment, many live in poverty and suffer from unequal healthcare access, primarily induced by social determinants (Földes and Covaci 2012; Cook et al. 2013; European Union Agency for Fundamental Rights 2018).

Figure 11. Official Romani flag. The Romani flag, accepted during the first World Romani Congress in 1971, has a bicoloured pattern representing the material world (green) and the heaven (blue). It also includes the wheel of a cart, both symbolising the migratory tradition of Romani and referring to their Indian origin by alluding to the dharmachakra symbol.

1.3.3 A genetic sketch of the Romani

To this point we have seen what is known about Romani purely from a historical and linguistic perspective, two sources which already strongly point at an Indian origin, one of the long-standing question/debates about this population, also given the misconceptions which kept haunting Romani since the "Little Egypt" (subsequently distorted in all European countries) affiliation label. As alluded previously in this paragraph, genetic studies contributed substantially to corroborate and clarify previous knowledge and provide new insights for the demographic history of the group, thus refining the level of complexity of previously inferred information and offering a biological basis for important claims about their history. Genetic research of the Romani started around the 1970s using blood markers and supporting the Indian origin, followed by founder mutations implicated in diseases (Bernasovský et al., 1976; Lasa et al., 1998; Piccolo et al., 1996b; Rex-Kiss et al., 1973; Sivakova, 1983). Studies kept focusing on the sharing of genetic variants with South Asian populations and, focusing on disease-causing markers[16], interesting discoveries highlighted that traits of biomedical relevance across multiple genes were specifically shared only with Indian/Pakistan individuals, thus reinforcing at least a strong connection with those regions (Abicht et al. 1999; Minárik et al. 2003; Morar et al. 2004; Azmanov et al. 2010). Subsequent studies, exploiting genome-wide data kept confirming a NW Indian origin due to higher sharing of variants, and thus similarities, with Indian population further corroborated by more identity-by-descent (IBD)[17] segments in common

[16] The first disorders highlighted to be shared with Indian populations involved a form of glaucoma involving a *LTBP2* gene mutation, another mutation linked to hearing loss associated with the *GJB2* gene and a condition causing altered neuromotor activity known as congenital myasthenia caused by a single base deletion in the *CHRNE* gene

[17] Identity by descent refers to a DNA sequence, shared by individuals, that was inherited from a common ancestor without the reshuffling action of recombination

with northern groups (Gujarati, Punjabi and Kashmiri) (Mendizabal et al. 2012; Moorjani, Patterson, et al. 2013; Melegh et al. 2017). SNP data revealed that South Asian ancestral contribution to the Romani's genomic pool is around one third of total ancestry (Moorjani, Patterson, et al. 2013; Font-Porterias et al. 2019), but the diversity within India is complex, being made of different components mainly represented by an ancestral north component (considered an ancestral western Eurasian component) and a southern one[18] (Reich et al. 2009; Moorjani, Patterson, et al. 2013; Moorjani, Thangaraj, et al. 2013). These results surely highlighted a connection with South Asia and particularly India, but also pointed out that other events must have acted to contribute to the remnant ancestral sources of Romani genome. To reinforce this "Indian connection" investigations on uniparentally transmitted markers, namely mitochondrial DNA (mtDNA), and Y chromosome data, also confirmed previous results. In the Y chromosome, paternally transmitted markers, such as the H1a-M82 haplogroup, can be found in north-western regions of India and also frequently in Romani groups (Pamjav et al. 2011; Zalán et al. 2011; Rai et al. 2012). As for the maternal side (mtDNA) different haplogroups shared between Romani and South Asian populations have also been reported to have originated in the latter geographic location. These haplogroups mainly belong to the M linage, such as M18, M25, M35 and M5a1b (Gresham et al. 2001; Gusmão et al. 2008; Mendizabal et al. 2011). Intriguingly, confirming the good quality of information extraction from genetic data of both SNP array data and uniparental markers, researches using the these two sources agreed on the period when the proto-Romani population would have left India, around 1,500-800 years ago, also corroborating what was

and thus identical to the ancestral copy. IBD can be used to measure the ancestral relationships among populations.

[18] Summarized as Ancestral North India, or Ancestral West Eurasian (ANI and AWE respectively) and Ancestral South Indian (ASI)

hypothesized based on historical data (Price 2000; Hancock 2002; Achim 2004; Mendizabal et al. 2012; Gómez-Carballa et al. 2013; Moorjani, Patterson, et al. 2013; Martínez-Cruz et al. 2016).

Even though the contribution of the European population to the Romani gene pool is quite well documented, there is not much data or consensus about what happened after the proto-Romani left India prior to their arrival in Europe (Ena et al. 2022). Nonetheless, uniparental haplogroups found in Romani and at high frequencies in Middle Eastern and Caucasian regions might help place a landmark in these regions for their diaspora (Derenko et al. 2013; Tarkhnishvili et al. 2014). However, genome-wide analyses on the putative Romani presence in these areas provided contradictory results (Mendizabal et al. 2012; Bánfai et al. 2018; Font-Porterias et al. 2019). The scenario gets clearer (or better studied) when Romani settled into the Balkans and West Eurasian ancestral component is considered. Indeed, their genomes retain ~80% of this ancestry (Moorjani, Patterson, et al. 2013; Bánfai et al. 2018). As a consequence of Romani differential gene flow within Europe, past patterns and relationships are still visible in extant genomes, as Romani show higher IBD sharing with groups living in eastern regions suggesting prolonged flow upon their arrival in the Balkans (Moorjani, Thangaraj, et al. 2013). The movements within Europe estimated using demographic modelling support the existence of the eastern and western main macro-groups of Romani, which diverged around 1000-900 years ago, with subsequent independent admixture events and reduction of population size (*N*e)[19] for the western group (Mendizabal et al. 2012). A generalized more abundant West European ancestry in Romani groups is commonly accepted, but its distribution has local differences. This

[19] The effective population size (*N*e) is a measure that represents the number of reproductive individuals of a hypothetical population (usually assuming Hardy-Weinberg equilibrium) presenting the same measured characteristic as assessed in an analysed population of interest.

ancestry source is indeed more of a patchwork of regional components rather than a static unbroken block of ancestry. This is reflected by the Balkan ancestry component contribution in current European Romani, where groups within the Balkan Peninsula retain higher ancestral proportion compared to north-western groups (Font-Porterias et al. 2019). Indeed, the authors showed that a more complex and refined pattern of ancestry sharing holds true for other regional components, where Romani residing in specific European regions tend to retain higher ancestral components from the local area than Romani residing elsewhere. Notably, a lower proportion of West Eurasian ancestry than previously thought (Moorjani, Patterson, et al. 2013) was calculated, because this ancestry can be overestimated due to the confounding effect of AWE (see footnote 18), a component representing admixture with West Eurasians occurred in South Asia prior to the beginning of Romani diaspora, thus independent from their migration history (Moorjani, Patterson, et al. 2013; Moorjani, Thangaraj, et al. 2013; Font-Porterias et al. 2019). This West Eurasian component in European Romani varies depending on the geographical location, with individuals form Balkans and central Europe displaying a lower (60%) proportion than others from more north-western parts like Baltic and Iberian regions (80%), thus reflecting the differing levels of admixture within Europe as a function of their northwest migration (Font-Porterias et al. 2019). As for the other regions of Romani diaspora, also uniparental lineages signal this West Eurasian admixture due to the presence of haplogroups specific of these locations in the genomes of nowadays Romani. Studies highlighted an inverse trend of decreasing typical South Asian uniparental lineages with distance form Balkans and detected West Eurasian haplogroups of both mtDNA and Y chromosome haplogroups (X, I-P259, J-M92, J-M67, T, U and H7, I1a, R, J1b3 respectively) corroborating genome wide data once again (Gresham et al.

2001; Kalaydjieva et al. 2005; Klarić et al. 2009; Pamjav et al. 2011; Zalán et al. 2011; Martínez-Cruz et al. 2016).

All these processes reconstructed so far played a substantial role in the creation of the extant variability of the Romani population, which likely were fostered by isolation and differential gene flow with host populations. Nevertheless, also major bottleneck events corresponding to key timepoints during their diaspora (out-of-India and out-of-Balkan) played a role in shaping their genetic landscape. When the proto-Romani population left India, probably in a single displacing event, it is estimated that a reduction of ~50% the original Indian *N*e took place (Mendizabal et al. 2012; Martínez-Cruz et al. 2016; Font-Porterias et al. 2019; Bianco et al. 2020). The second event, as anticipated above, occurred when Romani left the Balkan area and, during the process that gave rise to the Eastern and Western groups, Western European Romani reduced by ~30% their *N*e (Mendizabal et al. 2012). Had this latter process continued in a spiralling cycle of population shrinking and loss of genetic diversity (with consequent rise in potentially harmful variants), Romani would have suffered more severe consequences of a completely isolated population, but gene flow with non-Romani Europeans, probably fostered by forced "assimilation" policies, is recorded along the subsequent increase in the population size of both Eastern and Western groups (Fraser 1992; Bianco et al. 2020). Overall, reduced levels of diversity were identified for Romani groups in north-western Europe compared to eastern ones. This measure could fit in a scenario of migration and subsequent fragmentation of Romani towards northern and western regions (Mendizabal et al. 2012). On a broader scale instead, the experienced bottlenecks, endogamy practices and isolation events left identifiable traces in Romani genomes as a whole, when compared to "parent" Indian and "host" European populations (Bianco et al. 2020).

The complex history of the Romani population played a prominent role in shaping their extant genomic landscape: periods of isolation, reduction of population size and gene flow with neighbouring groups left specific mutational footprints that nowadays can be detected comparing them with other populations. Attention has been given to disease-causing mutations as the presence of high frequency or private variants of this kind were highlighted in Romani groups (Kalaydjieva et al. 1996; Kalaydjieva et al. 2001; Morar et al. 2004; Cabrera-Serrano et al. 2018; Font-Porterias, Giménez, et al. 2021) along with increased frequency of slightly deleterious variants and general reduction of diversity, signalling how past events of their history contributed to shape the current genomic make-up (Mendizabal et al. 2013). Given the Romani history summarized in this chapter, the existence of population-specific variants that undergone selective pressures could be interpreted as the result of forces that either acted when proto-Roma encountered different environments or controlled the frequency of deleterious variants. Even though signals of selection specific to Romani were not found, it is worth mentioning that the influence of the increased frequency of slightly deleterious variants has been mostly attenuated by admixture events (and thus gene flow) with the populations they interacted with (Dobon et al. 2020; Font-Porterias, Caro-Consuegra, et al. 2021). As stated previously, one of the indications that Romani originated from India was the existence of specific biomedically relevant mutations shared between these two populations, and the clinical screening of Romani further highlighted a number of conditions (metabolic syndrome, cardiovascular and respiratory diseases, overweight) that might account for the reported higher mortality and disease prevalence, even in light of contradictory results (Vozarova De Courten et al. 2003; Zeljko et al. 2008; Simko and Ginter 2010; Parekh and Rose 2011; Dobranici et al. 2012; Nunes et al. 2018; Werissa et al. 2019). Of course, medical and clinical genetics research are pivotal to elucidate patterns of inherited

health-related variants, but it is crucial to bear in mind that also socio-economic and cultural factors such as lack of close care centres and health insurance, problematic communication or direct discrimination (thus escaping specific biologically based investigations) can play an important role impacting Romani global health (Földes and Covaci 2012; Nunes et al. 2018). Overall, the higher frequency of deleterious mutations and incidence for specific diseases in Romani groups warrants a careful consideration of the potential contributing biological factors, even despite scarce medical evidence and particularly due to persistent discrimination and uneven access to healthcare compared to majority groups.

2. OBJECTIVES

The study of structural variations posed an interesting challenge to the way human population genetics has classically been approached by research, widely dominated by SNP-based studies for which specific analytic tools and methodologies were developed over decades, providing fine-grained levels of investigation. The discovery of this next level of complexity, previously unexplored, sparked the interest for the possible implications of genomic rearrangements that, due to their intrinsic characteristics of spanning more than a single base, might have had a larger impact than SNPs, or at least could have played a significant biological and evolutionary role. Increasingly complex and broad studies, addressing larger and more diverse datasets put a spotlight on the implications of SVs in diseases and biological processes, how different populations share common variants as much as they retain private and rare ones, the power of CNVs in highlighting patterns of variation among individuals and how selective forces influenced their frequencies within populations.

In this work, we analysed for the first time the copy number variation of Romani samples in Europe using whole genome sequences. In fact, even though the Romani population has been attracting the attention of researchers (also outside the biological field), whose work revealed different aspects of the complex history of migrations and admixture with other populations, extant analyses permanently relied on genetic variation in the form of single nucleotide changes. We then ventured in this pilot study aiming to describe the presence and type distribution of CNVs (primarily deletions and duplications) in Romani along with other reference populations from South Asian, Middle Eastern and European regions that, as previously described, had different contacts with Romani people during their diaspora. Addressing the population genetics of Romani using CNVs, the study design had indeed a two-tiered approach.

From one perspective, we studied CNVs in an underrepresented population that, despite being a minority ethnic group, has already been characterized quite well in terms of ancestral contacts and gene flow, demographic history through time and the presence of mutations of biomedical interest. Thus, bearing in mind what is the current knowledge so far, we used this situation as a case study to assess the ability of CNVs to recover a coherent history, either pointing to similar conclusions compared to previous results or not and, overall, evaluating the informative power of this type of genetic variation within an interesting population context.

The second aspect was somehow the specular reflection of the former, namely leveraging this unexplored marker type in such an underrepresented population, which already provided intriguing findings in terms of founder mutations (some implicated in biomedical features), scouting additional layers of information that, so far, escaped SNP-based research. Research in SVs proved how these markers have an influence in gene dosage levels, either disrupting transcribed sequences or their modulators and thus altering phenotypes with either evolutionary or biomedical consequences. In this context, such properties represent a thrilling opportunity in the framework of human evolution and population genetics investigation, even more so, in the regards of an isolated and underrepresented group.

To investigate the extent of CNV variation in Romani and related populations, we addressed the following objectives:

i. Establish a robust calling pipeline for CNV detection in a heterogeneous dataset.
ii. Reconstruct populations relationships using the retrieved CNVs, comparing the results with previous knowledge.
iii. Assess the informative power of deletions and duplications respectively.
iv. Scan for differential patterns of genome location of the variants among the studied populations.
v. Detect CNVs of possible biomedical interest, particularly focussing on a putative higher burden of deleterious variants in Romani.

3. RESULTS

Population history modulates the fitness effects of Copy Number Variation in the Roma

Marco Antinucci, David Comas, Francesc Calafell

Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain

Address correspondence to Francesc Calafell, francesc.calafell@upf.edu

Abstract

We provide the first whole genome Copy Number Variant (CNV) study addressing Roma, along with reference populations from South Asia, the Middle East and Europe. Using CNV caller software for short read sequences data we identified 3171 deletions and 489 duplications. Using deletions information, we managed to discern population structure and differences in the number/length of variants, confirming how CNVs carry and can disclose information about genetic variability among human populations. We estimated the amount of differentiation among population pairs, showing how Roma people share decreasing differentiation with closer Middle East and European populations, confirming previous knowledge. Roma history probably influenced patterns of genomic losses in their descendants, as we could highlight an increase in intronic deletions within Loss of Function (LoF)-intolerant genes. This points to slightly relaxed natural selection in Roma which affected mildly deleterious variants, though not those affecting exons. Over-representation analysis over LoF-intolerant gene sets hosting intronic variants highlights a substantial accumulation of shared biological processes in Roma, intriguingly related to signalling, nervous system and development features, which may be related to the known profile of private disease in the population. Finally, we show the link between deletions and known trait-related SNPs reported in the GWAS catalog, which exhibited even frequency distributions among the studied populations. This suggests that, in general human populations, the strong association between deletions and SNPs associated to biomedical conditions and traits could be widespread across continental populations, reflecting a common background of potentially disease/trait-related CNVs.

Introduction

Structural variants (SVs) are a class of genomic rearrangements, larger than 50 bp, comprising insertions, deletions, duplications, inversions and translocations, which are responsible for the largest fraction of base pair variation in the human genome (Sudmant, Rausch, et al., 2015; Weischenfeldt et al., 2013). Within SVs, balanced mutations (inversions and translocations) do not alter the genomic dosage, while unbalanced rearrangements (insertions, duplications and deletions, the latter two also known collectively as Copy Number Variants, CNVs) involve losses or gains of genetic material. CNVs can exert their influence on gene expression, phenotypic traits, and diseases, and represent a main source of genetic variation on which natural selection can act upon (Audano et al., 2019; R. L. Collins et al., 2020; Handsaker et al., 2015; Hollox et al., 2021; Hurles et al., 2008; Perry et al., 2008; Stranger et al., 2007). Indeed, CNVs have been linked to a number of traits such as Crohn's disease, osteoporosis, HIV susceptibility, body mass index, cancers and psoriasis (De Cid et al., 2009; Dentro et al., 2021; Hamdan & Ewing, 2022; McCarroll et al., 2008; Mohamad Isa et al., 2020; Willer et al., 2009; Yang et al., 2008) and are intriguingly associated to neurodevelopmental disorders in humans (Girirajan et al., 2013; Kato et al., 2022; Morris-Rosendahl & Crocq, 2020; Sebat, Lakshmi, Malhotra, Troge, Lese-Martin, Walsh, Yamrom, Yoon, Krasnitz, Kendall, Leotta, Pai, Zhang, Lee, Hicks, Spence, Lee, Puura, Lehtimäki, et al., 2007; Sekiguchi et al., 2020; Singh et al., 2017; Stefansson et al., 2008).

Most of the studies addressing human population genetics have historically focussed on SNPs to infer human population demography, such as changes in effective population size due to bottlenecks or founder events, or gene flow due to migration. This is also the case for the investigation of the mutation load, that is, the global contribution of deleterious mutations to

disease. However, research using CNVs as markers in population genetics surveys, both in large worldwide comparisons and on finer scales, has been increasingly accumulating over the last two decades and confirmed their potential in this field, highlighting among/within group variability, the functional potential of the variants (including pathogenic effects) and their evolutionary relevance (Almarri et al., 2020; Bergström et al., 2020; R. L. Collins et al., 2020; Dennis et al., 2017; Gautam et al., 2012; Hehir-Kwa et al., 2016a; Itsara et al., 2008; Redon et al., 2006; Sudmant, Rausch, et al., 2015; Urnikyte et al., 2016). We propose to use CNVs with these goals in a singular population, the Roma, which, as explained below, combine in their genomes the effects of gene flow, population subdivision and a reduction in effective population size.

The Romani or Roma population (often referred to by the problematic misnomer *Gypsies*) nowadays forms the largest transnational minority ethnic group in Europe; nevertheless, their origin has been traced back to North-western India thanks to different sources of information. The lack of self-written historical records makes it particularly difficult to portray a clear picture of their early history and interactions with surrounding groups. Still, linguistic studies and records from the populations that encountered the proto-Roma groups often suggest an Indian origin of this group, which left around 1000-1500 years ago and subsequently spread to Persia and Armenia (Boerger, 1984; Fraser, 1992; Liégeois, 1994). Records from Greece, present-day Romania, and the Czech Republic account for putative Roma presence in these territories through the 14th century, and by the 15-16th centuries, additional historical evidence documents Roma movements in many West European countries (Fraser, 1992; Liégeois, 1994). The current distribution of Roma people throughout Europe can be attributed to such early 15th century expansions from the Balkans and multiple dispersals in later times from the 19th century which were mainly triggered

by socio-economic reasons (Fraser, 1992; Gresham et al., 2001; Liégeois, 1994; Reyniers, 1995). In more recent historical times, the Roma population size and distribution in Europe is also the consequence of the genocide they suffered, carried out by the Nazi Germany regime (Lutz, 1995; Milton, 1991; Sridhar, 2006). Finally, the fall of the communist regimes in Central and Eastern Europe facilitated westward economically driven migrations.

The European Roma groups, indeed, have had a complex history, both in terms of the movements and contacts with different populations. Population genetics studies traced back their South Asian-related ancestry, with subsequent European admixture, from autosomal and uniparental markers (Font-Porterias et al., 2019; Gresham et al., 2001; Moorjani, Patterson, et al., 2013). Their specific history also shaped the landscape of genetic diseases, as different deleterious mutations were detected at higher frequencies, while other mutations are absent or at lower frequencies compared to other non-Roma populations (Kalaydjieva et al., 2001; Mendizabal et al., 2013; Morar et al., 2004). Specifically, private disease-causing mutations, highlighting a scenario typically found in a founder population, have been identified also in the Roma. The traits associated to these mutations are, among others, polycystic kidney disease, congenital glaucoma, congenital myasthenia, galactokinase deficiency, different neuropathies and centronuclear myopathy (Angelicheva et al., 1999; Cabrera-Serrano et al., 2018; Kalaydjieva et al., 1996, 1999, 2001; Morar et al., 2004; Piccolo et al., 1996a).

The whole-genome sequence of 46 Roma individuals revealed a strong, early founder effect followed by a drastic reduction of ~44% in effective population size (N_e) (Bianco et al., 2020). It is known that mutations reach fixation faster in small populations due to drift and, as a consequence, some deleterious mutations may rise in frequency and, under specific conditions,

slightly deleterious variants can result in a larger load than more deleterious ones (M. Kimura et al., 1963; M. Kimura & Ohta, 1969). In general, a rule of thumb is that drift will prevent the removal of deleterious mutations if $N_e s$ <1, where s is the selection coefficient; still, this does not encompass the complexities of population growth and gene flow (Gazave et al., 2014; Lohmueller, 2014). Different studies observed these phenomena in general populations as the Europeans, but also confirmed them in smaller and isolated groups which experienced more recent bottlenecks (i.e., Finnish, French-Canadians, Inuit and Ashkenazi Jewish) (Casals et al., 2013; Kaklamani et al., 2008; Lim et al., 2014; Lohmueller et al., 2008; Pedersen et al., 2017; Thaler et al., 2009). Moreover, disease-associated variants show specific haplotype ancestry backgrounds in Roma (European or South Asian), in line with the mutual contribution of these ancestries to Roma genetic makeup and, additionally, that the higher frequencies of SNPs mapping to drug-binding domains match the population higher proportion of diseases targeted by such drugs (Font-Porterias, Giménez, et al., 2021). This stresses how admixture dynamics, demographic history and the functional role of variants all contribute to the shaping of the extant diversity detectable nowadays in Roma.

In light of the information about Roma gathered so far, we hereby analyse for the first time CNVs in high-depth complete genomes from the underrepresented European Roma population to both confirm and provide new insights into their demographic history, as well as understand how this demographic history may have contributed (if at all) to their mutation spectrum and mutational load.

Methods

Samples

Our study comprises 40 complete genomes of Roma people collected in five European countries (Spain, Lithuania, Hungary, Ukraine and Macedonia) and belonging to four major migrant groups: 15 North/Western, 5 Vlax, 10 Romungro and 10 Balkan as defined in a previous study (Bianco et al., 2020). Donors signed an informed consent and the project was approved by the Institutional Review Board of the Comitè Ètic d'Investigació Clínica-Institut Municipal d'Assistència Sanitària (CEIC-IMAS) in Barcelona, Spain, (2016/6,723/I). All participants self-identified as Roma and appropriate consent was obtained from all donors. The study was approved by our IRB (Comitè d'Ètica de la Investigació, Parc de Salut Mar, Barcelona) on June 7th 2016 (reference 2016/6723/I) and renewed on January 15th, 2020 (reference 2019/8900/I). Preliminary results were presented to the Roma community in a meeting on February 1st 2019 in Barcelona. All methods in this study were performed following the standard guidelines and regulations. Genome sequences were those analysed in (Bianco et al., 2020), which fastq files had been deposited at the European Genome Archive with accession number EGAS00001004287. Reference samples with geographic origins matching the Roma diaspora comprised two main datasets: the Simons Genome Diversity Project (SGDP; samples from Europe, the Middle East and South Asia) (Mallick et al., 2016) and Mondal et al. (Mondal et al., 2016) (samples from South Asia).

Structural variant calling

We selected a set of six different programmes using algorithms based on different strategies to detect SVs from short read sequencing data, combining the strengths of each algorithm and integrating them. Our set is composed of CNVnator (version 0.4.1) (Abyzov et al., 2011), BreakDancer (version 1.4.5) (K. Chen et al., 2009), Pindel (version 0.2.5b8) (Ye et al., 2009), Tardis (version 1.0.4) (Soylev et al., 2017), Lumpy (version 0.2.13) (Layer et al., 2014), and GenomeSTRiP (version 2.0) (Handsaker et al., 2011, 2015) callers, which implement read-depth, split-read and read-pair methods. See Supplementary Methods for the implementation of each method.

Data merging

We designed custom scripts to obtain the data both for the results for all callers for a single sample and among all samples. To do so, we first merged the output of the different software for each sample, specifically by merging those SVs residing on the same chromosome, deletions and duplications separately, with a reciprocal genomic coordinate overlap of at least 50% of their length. By doing so, we created clusters of overlapping pairs of calls and for each cluster (ranging from a pair of calls for two programmes, up to 15 pairs, corresponding to all possible combination of pairs –without self-pairing- among the six software used) we selected the coordinates and the genotype of the most confident caller, based on the evaluation of caller performance in (Kosugi et al., 2019). Using this information for each cluster of calls mentioned above, we obtained a single call by retaining the best performing software for coordinates and genotype respectively. To merge variants across samples we proceeded in a similar manner as previously presented, where we joined all sample calls if variants of the same type resided on the same chromosome and reciprocally overlapped at least for

50% of their length. This allowed us to create a consensus set of calls listing the sharing of each variant among individuals.

Additional filters

We regenotyped the CNVs of each sample with a dedicated software, GraphTyper2 (version 2.5.1) (Eggertsson et al., 2019), to accurately recover more reliable genotypic information. We further filtered the results according to the best practices as described by the software authors, to retain only good quality genotypes. To additionally filter for false positives, we used the HardyWeinberg R package (version 1.7.2) (Graffelman, 2015) to remove variants violating Hardy-Weinberg equilibrium. We computed the chi-squared test p-value for each CNV in each population and filtered out variants having a significant result after Bonferroni correction for multiple tests. Finally, we implemented an R package algorithm leveraging SNP data to infer reliable CNVs: CNVfilteR (version 1.8.0), which detects false positive heterozygous deletions and duplications by evaluating the frequencies of SNPs mapping to each variant (Moreno-Cabrera et al., 2021). We ran this software with default parameters and obtained a set of variants indicating false positive results that were subsequently filtered out from the dataset.

Statistical analysis

Principal component analysis was carried out using the smartpca algorithm within the Eigensoft package (version 6.0.1) (N. Patterson et al., 2006). Briefly, based on CNV genotypic calls, we coded biallelic deletions and duplications as zero, one, and two copy numbers and used those as input for the software to perform PCA on our samples. We additionally used another dimensionality reduction method, the uniform manifold approximation projection (UMAP) (McInnes et al., 2018) on copy number for deletions and duplications. Population structure was further assessed

using ADMIXTURE (version 1.3) (Alexander et al., 2009), running 10 random seeds for each ancestral component (K: 2 to 10), to evaluate ancestry profiles among the studied samples. We filtered out variants with minor allele frequency < 0.01 and violating structure-aware Hardy-Weinberg equilibrium before running the analysis, as best practices described in previous studies (Hao & Storey, 2019; Linck & Battey, 2019; Narang et al., 2014). Pong (Behr et al., 2016) was used to visualize ADMIXTURE results by representing Q matrices for modes in each value of K. ANOVA test was performed with the R *car* package (version 3.0.10) (Fox, John & Weisberg, 2011), while Kruskal-Wallis and Chi-squared tests were computed using the corresponding native R functions (R Development Core Team, 2003). We estimated global differentiation values calculating F_{ST} statistics among pairwise populations using the StAMPP R package (Pembleton et al., 2013) and estimated p-values by performing 10000 bootstraps. Taking advantage of the possibility to recapitulate population differentiation using CNVs data by means of the Vst statistic (Redon et al., 2006; Sudmant, Mallick, et al., 2015), using a custom script, we implemented a variation of the formula described in a previous study (Serres-Armero et al. 2021), comparing directly copy number variance rather than log2 ratios from CGH array data. We applied the statistic in pairwise population comparisons computing the differentiation for each CNV individually.

Copy number variant annotation

We used the software AnnotSV (version 3.0.7) (Geoffroy et al., 2018, 2021) for multiple database annotation to retrieve the possible clinical or functional roles of the CNVs in our dataset. Since results from AnnotSV provided different information, we focussed on: 1) the genes intersected by the CNV, 2) whether the intersection involved an intron, an exon, or both, 3) diseases associated to the intersected gene provided by OMIM catalogue

(Hamosh et al., 2005), 4) overlap with Topologically Associated Domains (TADs), 5) gene tolerance to loss of function. Specifically, the tolerance to loss of function for genes intersected by CNVs is ranked as LOEUF (Loss-of-function Observed/Expected Upper Fraction) bins (range 0 to 9) from genomAD database (Karczewski et al., 2020). The LOEUF metric refines over the widely used pLI (probability of Loss of function Intolerance), providing a continuous rather than a dichotomous scale (e.g pLI < 0.9; pLI > 0.9). We carried out permutation tests to screen for possible intra-population higher/lower than expected abundance of deletions intersecting intronic portions of loss of function (LoF) intolerant genes. To do so, we downloaded the LOEUF information for each gene present in the gnomAD database and obtained those genes' annotations via Ensembl database (version 86) (Cunningham et al., 2022) using the EnsDb.Hsapiens.v86 and ensembldb R packages (Rainer, 2017; Rainer et al., 2019). For this list of genes, we extracted the intronic coordinates using GenomicFeatures R package (Lawrence et al., 2013) of those genes with a LOEUF <= 4 (Lof intolerant) and LOEUF > 4 or not reported (LoF tolerant). Then, with our list of population-specific gene-intersecting deletions and introns coordinates of LoF tolerant/intolerant genes, we performed permutation tests separately in each population using the regioneR R package (Gel et al., 2016) performing 5000 permutations and estimating the numOverlaps and randomizeRegions as the evaluate and randomize functions.

Over-representation analysis

To assess putative significant enrichment in biological pathways for our gene-intersecting SVs, we interrogated the Gene Ontology Resource (Ashburner et al., 2000) using the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) (Liao et al., 2019a; B. Zhang et al., 2005), an online tool to interpret and analyse gene lists of specific interest. We tested whether the list of genes classified with a LOEUF score from 0 to 4 and hosting intronic

variants was enriched in specific GO terms in each population. Accordingly, the inputs passed to the software were the above mentioned gene list as well as a reference set, namely all genes (regardless of their known intolerance level) having intronic deletions. We focused our analysis on biological and molecular function database categories, performing the analysis with default parameters and considering as significant the associations having an FDR < 0.05.

CNVs and GWAS catalog

We evaluated the level of association between our set of CNVs and diseases identified in the GWAS catalog (Buniello et al., 2019), using linkage disequilibrium (LD) with trait-associated SNPs as a proxy. As described by (Valls-Margarit et al., 2022), LD between CNVs and SNPs can be confidently estimated, and our analysis was based on SNPs shared between our dataset and the GWAS catalog. The selected common variants underwent filtering using PLINK (version 1.9; www.cog-genomics.org/plink/1.9/) (Chang et al., 2015), removing individuals with a missing genotype rate > 0.1 and SNPs with missing call rate > 0.1, with minor allele frequency < 0.01 and those failing the Hardy-Weinberg equilibrium test. This set of filtered SNPs and our CNV set were merged together and phased using two programs, WhatsHap (version 1.1) (M. D. Patterson et al., 2015) and ShapeIt4 (version 4.1.3) (Delaneau et al., 2019), following procedures previously described (Valls-Margarit et al., 2022). The result provided the input for PLINK, where we computed LD between variants in our dataset (CNVs and SNPs) and those SNPs shared with the GWAS catalogue, only including variants in high LD (r^2 > 0.8) and mapping within 1 MB around the pathogenic SNP.

Results

Calling CNVs from whole genome sequences

We called CNVs in 40 genomes from already published Roma individuals (Bianco et al., 2020; García-Fernández et al., 2020) along with 98 samples from Europe, the Middle East and South Asia (Mallick et al., 2016; Mondal et al., 2016). Our calling pipeline comprised six programs (callers) for SV detection from WGS using hg38 as reference genome; Table 1 recapitulates the average variant count for each caller and type of CNV, prior to the filtering before merging (see Methods).

For our subsequent analyses, we included only deletions and duplications (DELs, DUPs) as some of the software used are unable to call insertions or inversions. We merged our data together by, first, creating a per-sample consensus among callers, finding 1484 ± 366 CNVs per sample on average (deletions: 1433 ± 352; duplications: 51 ± 23) and eventually by iteratively merging sample CNVs, obtaining calls for individuals sharing the same variant (see Methods). This step yielded a total number of 11207 CNVs (9863 deletions and 1344 duplications) and an average of 1499 ± 352 CNVs per genome (deletions: 1449 ± 357; duplications: 50 ± 22).

Dataset characteristics and population structure

We grouped our 138 samples using a geographical rationale and divided the samples as follows: Roma (40 samples), Europe (22), Middle East (15), and South Asia (61). Initially, Principal Component Analysis (PCA) revealed that samples clustered by dataset of origin (Roma, Mondal et al. (2016) and SGDP) rather than by geographic affiliation (Supplementary Figure 1). This reflects batch effects, which, in this context, can be described as differences in the dataset that are caused by technical issues rather than by any underlying biological processes. Batch effects have been described also

when using CNVs; indeed, differences in sample preparation (PCR vs PCR free) and sequencing (insert size, read length, coverage) can introduce intrinsic features in each dataset that will affect the calling step and will account for the major differences in PCA (Almarri et al., 2020; R. L. Collins et al., 2020). We addressed this batch effects following a two-step procedure: first we regenotyped each CNV call to obtain more accurate deletions and duplications genotypes; second, we filtered out putative false positive results using jointly two strategies (see Supplementary Methods).

We performed PCA on our regenotyped and filtered dataset. PC1 and PC2 in both deletions and duplications still recover variance due to dataset-specific features (Supplementary Figure 2 and 3). However, deletions show a clearer structure in PC3 and PC4, where Roma tend to cluster between Europeans and South Asians, as expected by their known degree of admixture with these populations (Font-Porterias et al., 2019). Additional visualization of the data was achieved using the uniform manifold approximation projection (UMAP) (McInnes et al., 2018) on copy numbers, showing consistent sample apportionment with PCA results and providing clearer structure (Fig. 1). We confirmed this result by performing admixture analysis in our dataset, showing a decreasing gradient of West Eurasian ancestry from European to Southeastern samples and the presence, at K3, of a specific Roma component in the dataset, which is also residually found elsewhere (Supplementary Figure 4). Intriguingly, a part of the samples in the South Asia population is from Pakistan and shows higher rates of West Eurasian component compared to the rest of the South Asian individuals, in line with previous reports (Laso-Jadart et al., 2017; Shan et al., 2021). We attenuated batch effects by removing false positives variants and re-assessing incorrectly genotyped calls in our initial set of CNVs, by doing so, we noted improvements in different data visualization methods and concordant results from the admixture analysis.

Out of 3660 CNVs, 1899 (52%), 329 (9%) and 459 (13%) are shared by four, three and two populations respectively. We additionally found 973 (27%) variants that were found in only one population (Roma: 257, Europe: 157, Middle East: 179 and South Asia: 380), most of which were singletons. Overall, our call set is composed of 2013 common (Allele Frequency, AF) > 0.05), 668 low frequency (0.01 <= AF <= 0.05) and 979 rare variants (AF < 0.01). Most common variants are shared preferentially by all four populations (four populations: 1792 (89%), three populations: 120 (6%), two populations 78 (4%), one population 23 (1%)) as expected in general populations. Low frequency variants are more evenly distributed (four populations: 107 (16%), three populations: 209 (31%), two populations 237 (36%), one population 115 (17%)) while rare variants, as expected, can be found only in one population or two at most (two populations: 144 (15%), one population: 835 (85%)). Within-population proportions of common, low frequency and rare variants change across populations, with South Asians having more variants across the frequency classes compared to the other populations and the Roma showing the same trend compared to Europe and Middle East ($\chi^2 = 83.6$, p-value = 6.25×10^{-16}) (Table 2). Globally, South Asia and Roma retain a higher number of private CNVs and, evaluating the frequency profiles among populations, this pattern repeats within common, low-frequency and rare variant classes, demonstrating that the apportionment of private variants is not restricted to any specific frequency category.

CNV annotation

Using the software AnnotSV (Geoffroy et al., 2018, 2021) we annotated variants leveraging different databases (Refseq, OMIM, ClinGen, gnomAD, among others) and gathered information about CNV localization within genes, their possible functional role and the pathogenic consequences of their presence in transcribed genome sequences. While

more than half of the CNVs in our dataset, 2115 (58%), did not overlap any currently known gene, 1532 (42%) variants intersected transcribed sequences, of which 263 (7.2%) and 1268 (35%) resided within exons and introns respectively, in agreement with previous studies (Conrad et al., 2010; Mills et al., 2011; Valls-Margarit et al., 2022). The remaining 13 CNVs intersected more than one gene, hitting multiple intronic and/or exonic locations. Overall, we found that genomic location and the type of CNV are dependent from each other (χ^2= 77.3, p-value < 2.2×10^{-16}), with deletions representing the majority of variants within each genomic location (Table 3). It is interesting to notice that exons seem to tolerate duplications better than deletions: while 6.7% of deletions affect exons, this figure is 18.2% for duplications, likely due to the stronger selective constraints over deletions within genes (Sudmant, Mallick, et al., 2015). Our dataset confirms what previous studies reported about the average frequency apportionment of intergenic and genic variants and the easier-to-resolve deletion signal used by short reads structural variants software.

Geographic and genomic distribution of CNVs

We next tested for the number and length of CNVs carried by individual. Statistical tests for deletions and duplications produced significant results mainly for the first category. For duplications, we could not find any relevant significant differences among the populations. As for deletions, Roma carry more events per individual (mean: 880 ± 24) with respect to all other populations, (Europe: 834 ± 16; Middle East: 828 ± 29; South Asia: 810 ± 26), (Anova p-value < $2.2x10^{-16}$). Testing for deletion location, we found out that the same pattern held true for intergenic (Kruskall-Wallis p-value < $2.2x10^{-16}$) and intronic (Kruskall-Wallis p-value = 10^{-14}) events (Fig. 2). Regarding exonic deletions, Europeans carry significantly fewer variants compared to Roma, Middle East and South Asia populations (Anova, p-value = 0.007). In addition, variant length also differed among

populations as, overall, deletions in the Roma are larger than those in Europeans, while deletions in South Asians are shorter compared to all other populations (Kruskall-Wallis p-value = $1.3x10^{-8}$) (Supplementary Figure 5A). In particular, Roma have larger variants only when considering intergenic deletions, while South Asian population shows shorter intergenic, intronic and exonic (Kruskall-Wallis p-values, intergenic = $8.2x10^{-10}$; exonic = 0.0016 and ANOVA p-value intronic = 0.0001), (Supplementary Figure 5). Overall, the results of these first comparisons show that Roma carry more and longer intergenic deletions than other populations, but their intolerance to exonic deletions is similar.

CNV differentiation among populations

Overall, the average F_{ST} (Figure 14) among all pairs of populations was higher for deletions (0.0375) than for duplications (0.0272), which is consistent with repeat mutation at duplications counterbalancing population differentiation by drift. Thus, we will base our population inferences on deletions. The average F_{ST} between the Roma and each of the other populations was 0.0478, which is higher than for any other population. In particular, the Roma were slightly more distant from South Asia (0.0497) than from the Middle East (0.0473) or Europe (0.0465). South Asia is also equally distant from the Middle East (0.0363) and Europe (0.0383), while these two populations are close to each other (0.0067). This is the expected pattern as derived from nucleotide variation in arrays (Granot et al., 2016) or whole genomes (Mallick et al., 2016). Particularly for the Roma, these differentiation patterns are in line with previous studies based on genome-wide SNP data (Melegh et al., 2017) and could reflect the global landscape of CNVs in Roma, who had their own mutational history diverging from Northern India, ultimately admixing with Europeans and, in the process, accumulating genetic drift.

Features of the highly differentiated CNVs

Next, we characterized the CNVs that were highly differentiated among populations by computing the V_{ST} statistic (Redon et al., 2006) for each CNV and pair of populations. Unlike F_{ST}, V_{ST} takes into account the variance of copy numbers in pairwise group comparisons. The mean V_{ST} values are reported in Supplementary Table 1. We focused on highly differentiated CNVs by taking the top 20% V_{ST} values, for each pair of populations (Figure 15); the average V_{ST} values by genomic location in this highly differentiated set can be found in Supplementary Table 2. Intergenic deletions and duplications are at the top of the value distribution; indeed, as expected, these variants display fewer constraints in the mutation rates between populations and thus are freer to vary. Intronic and exonic variants follow in the distribution, showing lower values for the latter calls and pointing once again to a higher constraint on those deletions and duplications putatively having a higher disruptive power over genic sequences. Since pairs containing Roma exhibited higher values at the top of the distribution, we tested if any difference existed in V_{ST} values among pairs for variants intersecting genes. We found significant differences (Kruskall-Wallis, p-value < $2.2x10^{-16}$) for deletions in such pairs with respect to the others. In particular, pairs considering Roma had significantly higher values than pairs without and, dividing the analysis by variant location, we could find significant differences only for intronic events (Kruskall-Wallis, p-value < $2.2x10^{-16}$; mean values: Roma-Europe = 0.1316; Roma-Middle East = 0.1452; Roma-South Asia = 0.131; Europe-Middle East = 0.0952; Europe-South Asia = 0.0894; Middle East-South Asia = 0.0923). Estimating variant differentiation among pairs of populations highlighted how the major source of variability can be traced back to Roma individuals, nevertheless, when stratifying the analysis by

genomic location of the variants, significant differences in differentiation scores can solely be found for intronic deletions.

Predicting the pathogenicity of CNVs

For each CNV we retrieved, whenever available, the OMIM (Online Mendelian Inheritance in Man) annotations (Hamosh et al., 2005) and the LOEUF (Loss-of-function Observed/Expected Upper Fraction) bin values (ranging in bins from 0 to 9) from gnomAD (Karczewski et al., 2020) when the variant overlapped a gene sequence. On one hand, OMIM entries refer to known disorders linked to specific phenotypes and/or genes, providing information about the putative deleterious role of variants at pathogenic genes. On the other hand, LOEUF indicate the degree of intolerance to loss of function (LoF) for a gene, suggesting the likelihood for a variant to disrupt gene function. We compared the distribution of variants hitting genes having a linked OMIM entry among populations and Europeans showed a significantly lower number (Anova, p-value = 0.02) of deletions within OMIM genes compared to all other populations (mean DELs per genome: Roma = 87.8, Europe = 82.5, Middle East = 86.4, South Asia = 85.5). Duplications, instead, are significantly (Kruskall-Wallis, p-value = 0.007) more frequent in South Asians than in Roma and Middle East populations (average of 6.85, 5.75 and 5.4 duplications per genome respectively). These results could highlight a greater efficacy in Europeans of natural selection removing deleterious mutations, probably due to their demographic history. Duplication within OMIM genes being more frequent in the South Asia compared to Roma and Middle East populations could reflect, to a certain degree, the increased recessive diseases specific to the group and the different selective pressures recorded for specific West Eurasian alleles, as highlighted in (Ayub & Tyler-Smith, 2009; Nakatsuka et al., 2017). To further assess the putative pathogenic effect of deletions, we summarized the number of variants of every LOEUF bin in each

population and tested the presence of differences among all populations for the most intolerant categories (0-4). Roma individuals showed increased number of deletions in the 0, 1, 2 and 4 bins and, upon stratification by location, only intronic events produced significant results for the same categories (bin 0: Anova, p-value = $2.8x10^{-7}$; bin 1: Kruskall-Wallis, p-value = $1.4x10^{-8}$; bin 2: Anova, p-value = $3.5x10^{-6}$; bin 4: Anova, p-value = $1.6x10^{-5}$). We assessed whether this higher number of deletions intersecting genes with low LOEUF values caused the overall increased number of intronic variants in Roma, as shown above. After removing these intolerant-gene deletions, Roma keep retaining a significantly higher number of intronic variants (Kruskall-Wallis, p-value = $5.2x10^{-10}$), demonstrating that the accumulation of these deletions at intolerant genes is an independent process that does not drive the general increase in intronic deletions. Due to our findings of an increased number of deletions within introns of LoF intolerant genes in Roma, we explored, separately for each population, the possibility that these mutations preferentially hit intronic coordinates while taking into account LoF tolerance. Permutation tests were performed using all genic deletions against intronic coordinates of genes either with a LOEUF <= 4 (intolerant) or LOEUF > 4 - or for which the metrics was not available - (tolerant). With these sets of regions we noticed that, while genic deletions intersect introns of tolerant genes more often than expected by chance (Permutation test, p-value = 0.0018-0.0004), the opposite is not true for the intersection with introns of intolerant genes (Permutation test, p-value > 0.05). This result points toward a general constraint for the accumulation of deletions, even at the intronic level, in intolerant genes within each population. In the context of the most differentiated variants described above, we looked at the distribution of frequencies and LOEUF values in pairwise populations containing Roma; we evaluated the frequencies in deletions showing larger differentiation, partitioning the variants across the most intolerant LOEUF classes (0-4). Despite the fact

that the only significant result showed higher frequency in Roma compared to Middle Eastern population for deletions in the LOEUF 2 category (Kruskall-Wallis, p-value = 0.02), we noticed a general trend towards slightly higher frequencies in the Roma, across all LOEUF bins, compared to all other populations (Kruskall-Wallis, p-value = 0.0471). Nonetheless, pairwise group comparisons do not show significant results after multiple test correction. Following our previous results on the differentiation of intronic deletions in Roma, here we show an over-representation of such variants in this population that, together, highlight a pattern of recurring mutations occurring in untranslated genome portions. The differences in intolerant-gene deletions could highlight a lower constraint for Roma towards the accumulation of genic deletions residing outside the coding sequences but within genes whose function is more likely hampered by mutations.

CNVs and Topologically Associated Domains (TADs)

Next, we annotated CNVs intersecting Topologically Associated Domains (TADs). TADs are defined as genome portions in close physical contact due to the three-dimensional arrangement of DNA sequences, which are more likely to interact with one another than with adjacent sequences. We assessed the degree of overlap between CNVs and TADs and identified eight variants (five deletions and three duplications, Supplementary Table 3), of which two singletons, intersecting these regions and at least one gene. Among our findings, we noticed interesting examples of deletions at clinically relevant genes, such as a 183kb deletion which affected both the first intron of *ADGRL4*, a gene involved in angiogenesis, and a TAD (coordinates: 1:79254316-80254315) found in the initial portion of the gene. Another deletion completely removes a gene involved in steroid metabolism, *UGT2B28*, and part of a TAD sequence encompassing other related genes (*UGT2B4* and *UGT2A2*). Furthermore, an intriguing example

involves a 151kb deletion spanning five immune system-related genes (*IGHG4*, *IGHG2*, *IGHA1*, *IGHG1* and *IGHG3*) on the telomeric portion of chromosome 14. The three most distal genes are also contained in an extended TAD sequence (coordinates: 14:105653664-106053818) including a dense cluster of immunoglobulin genes. In these examples, the frequency of the deletions is evenly distributed among the populations sharing the variant, except for the latter, where Roma show the highest (0.313) within an increasing frequency cline from South Asia (0.107) to Middle East (0.167) to Europe (0.205). It is also worth noticing that all the TAD-intersecting variants share a same common feature: their relative large size. Indeed, the five deletions have a mean size of 203kb (compare to the overall 4kb mean for deletions), while duplication average length is 193kb (as opposed to their 9.4kb mean size across the genome). This may be just a spurious effect: longer CNVs may be more likely to intersect TADs or any other genomic feature just because they are longer.

Duplications are either singletons/rare events (such as a duplicated sequence overlapping *GRHL2* transcription factor in 3 individuals), or widespread (present in 78 samples mainly intersecting *NBPF1* a member of a highly duplicated family (Vandepoele et al., 2005)). Together, these findings provide hints for 3D genome conformations whose function - typically fostering gene transcription by facilitating the action of enhancers upon gene promoters (Beagan & Phillips-Cremins, 2020) - might be impaired or disrupted by deletion events, particularly when multiple genes encoding similar products reside closer together (Lupiáñez et al., 2016; Shanta et al., 2020).

CNVs and genetic associations

In Genome Wide Association Studies (GWAS), genetic associations are established between specific diseases or traits, or sets of them, and genetic variants, usually SNPs. We wondered to which extent the CNVs we detected could be linked to pathogenic SNPs present in the GWAS catalog (Buniello et al., 2019). To do so, we downloaded the GWAS catalog dataset version 1.0.3 and identified common SNPs between this set and those previously found in our samples (Bianco et al., 2020); the intersection consisted of 74,009 variants. For these common SNPs, we estimated the associated CNVs by selecting, for each chromosome, only those CNVs in strong linkage disequilibrium (LD) ($r^2 > 0.8$) and residing in a 1 MB window around the SNP. Following this procedure, we identified 78 unique deletions in LD (supplementary Table 4) with 125 disease-associated SNPs as reported in the GWAS catalog, while no duplication was in linkage disequilibrium with any SNP in the set. The identified deletions are in LD with one or more (up to eight) SNPs and, for each of them, we retrieved the information about deleteriousness using LOEUF scores. Among the traits in the GWAS catalog, we could identify different functional categories. The majority of the traits involves metabolic, neurodevelopmental/neurological, development and haematological-cardiovascular disorders. Looking at the genomic context of the linked deletions, 41 (53%) reside in intergenic loci, 32 (41%) intersect introns and only five (6%) within exons. While a direct role of intergenic variants upon the pathogenicity of linked SNPs is difficult to establish – but not a reason to exclude them *a priori* – intronic and exonic CNVs might act on the same genomic context of the SNP. Among the intronic variants, only eight deletions intersected genes having more tolerant LOEUF scores (> 5), six other gene-intersecting variants had no score information and the remaining 18 resided in genes with higher intolerance to LoF (scores 0-4). Among these latter deletions, four are in

linkage with SNPs related to metabolic/inflammatory diseases (Type 2 diabetes, alanine transaminase levels, urate levels), four others link with GWAS traits related to heart, cardiovascular or haematological conditions (myocardial infarction, haemorrhoidal disease, red-cell width) and two variants link to colorectal cancer traits. For exonic variants, only one deletion intersects an intolerant gene (LOEUF bin 4) and is in LD with a SNP associated to metabolic disorders (total cholesterol/LDL levels); nevertheless, the deletion resides in a gene upstream the SNP and its involvement is unclear. The remaining four exonic deletions associate with inflammatory diseases, lung function, haematological and developmental features and all but one (lung function) affect the same gene of the linked SNP. Nonetheless, intolerance scores are either not available or point to a relaxation against LoF for exonic variants. Finally, when considering only the set of SNPs residing ~5000bp around linked deletions, we noticed that intergenic events are the most frequent type of variants in the set (19 intergenic deletions, against nine intronic and one exonic deletions). This evidence, at least in part, might support the hypothesis of a possible influence, due to physical proximity (71bp for the closest intergenic deletion), upon the genomic environment shared with the associated pathogenic SNP. In general, using data from the GWAS catalogue, we were able to leverage SNPs information as a proxy for putative CNVs involvement in health-related traits, showing that either co-occurrence of a deletion and a SNP within the same gene or physical proximity may add novel information to both the traits and to the function of the structural variant under investigation.

Functions of the genes affected by deletions in the Roma

As previously shown, our analysis on deletion pathogenicity showed that the Roma retain a higher number of deletions intersecting LoF intolerant genes, and that specifically intronic variants are responsible for this result. With this observation at hand, we wondered whether these more abundant intronic deletions in Roma had a specific influence on biological processes. We tested this hypothesis by performing an over-representation analysis separately in each population, using the online software GEne SeT AnaLysis Toolkit (WebGestalt) (Liao et al., 2019b; B. Zhang et al., 2005), assessing whether LoF intolerant genes (LOEUF bins: 0-4) intersected by intronic deletions were present more than expected in Gene Ontology (Ashburner et al., 2000) terms (GO terms). Results show significant enrichments in GO terms for the set of input genes in each population, with a marked prevalence of associations in Roma. Indeed, while Europe, Middle East and South Asia populations had significant enrichment for 24, 18 and 37 GO terms respectively, Roma's significant GO terms amounted to 187. For each term, using the available descriptions of related biological processes, we identified three recurrent functional categories, namely Nervous System, Signalling and Development, plus a catch-all *Other* category (Figure 16). Overall, Roma showed higher number of GO terms among these classes compared to reference populations. The two most abundant categories in Roma were Signalling and Nervous System, which contained 61 and 55 GO terms respectively. As a comparison, these two categories included 13/0, 0/2 and 6/11 terms in Europe, Middle East and South Asia respectively. Furthermore, using a function within WebGestalt aiming at reducing possible redundancy for GO terms having similar gene sets, we obtained clusters of terms sharing related biological processes. Following this clusterization, the Roma had 33 GO clusters, including 11 Signalling, 9 Nervous System, 5 Development and 8 comprising other

processes such as chemotaxis, cell motility and cellular component organization. Europe, Middle East and South Asia had five, four and eight clusters with different proportions of the three major functional categories. We additionally checked for significant GO terms specifically found only in one population and noticed that Roma retain the highest number of private significant results, with 125 private terms against three, one and six found in Europeans, Middle Eastern and South Asian samples. Considering the deletions intersecting genes associated to the 125 private GO terms in Roma, we obtained 410 variants and retained only those overlapping a known pathogenic gene, either annotated in the OMIM or Deciphering Developmental Disorders (Firth & Wright, 2011) (DDD) databases. The final filtered set included 168 deletions whose frequencies do not vary noticeably across populations; nevertheless, it is interesting to highlight that out of the 23 rare deletions, considering the global frequency in the whole dataset, 21 are indeed private to Roma. Within these Roma private deletions, more than half (15 variants) are singletons and reside in genes mainly associated to developmental/neurodevelopmental diseases and cancer. Of the remainder six deletions, four are doubletons and reside in genes associated to Cerebellofaciodental, Bardet-Biedl, Gillespie's syndromes, spinocerebellar ataxia 15 and skeletal dysplasia with severe neurological disease, while the two more common variants intersect genes linked to Phelan-McDermid syndrome and -2-hydroxyglutaric aciduria. Overall, further investigation of intronic deletions in LoF intolerant genes revealed significant enrichment in biological processes mainly related to signalling, nervous system and development, with a sharp accumulation of GO terms in Roma compared to the other populations. This supports our results of higher differentiation and abundance of intronic deletions within Roma, suggesting a possible relevance upon the functions of genes sets bearing such variants.

Discussion

In the current study, we analysed CNVs in the Roma population using whole genome sequencing data with the dual purpose to provide the first published catalog of genome-wide unbalanced structural variants and, given previous knowledge of Roma demographic and genetic history, assess to which extent CNVs can inform us when used in a population genetics study of an underrepresented community. Comparing deletions and duplications from Roma and other reference populations (samples from Europe, Middle East and South Asia, covering the dispersal route Roma crossed in their diaspora) we estimated the main differences in the apportionment of events, the differentiation among populations and assessed the potential biomedical impact of the variants.

Removal of batch effects

We used genome sequences from three sources: Roma (Bianco et al., 2020), South Asians (Mondal et al., 2016) and European/Middle Eastern/South Asian (Mallick et al., 2016) sequences which, via PCA using genotyped copy numbers, maintained a three clusters structure regardless the overlap in geographic groups. Thus, we were detecting a batch effect. Batch effects had been previously described for CNVs (Scharpf et al., 2011) and, more recently, continued suffering from differences in library preparation, coverage and from the issues mentioned above concerning whole genome sequencing (Almarri et al., 2020; R. L. Collins et al., 2020).

By applying our filters described above, we removed ~67% of the initially identified variants in the whole dataset and further inspections using PCA and UMAP approaches for dimensionality reduction showed improvement of the clustering among populations, highlighting a more homogeneous geographic-wise distribution of the samples. We noticed that the two first components in PCA still retain some separation among the three datasets,

even though the severe clustering seen in the unfiltered dataset is no longer detectable, while PC3 and PC4, despite presenting some noise, clearly are not affected by batch effects. UMAP result are in agreement with this view, by showing the expected population clustering and closer mapping of samples from different datasets. We also show that different dimensionality reductions methods produce different results; PCA uses a linear transformation on the data and can be influenced by outliers. UMAP, on the other hand, is a non-linear method and performs better at preserving local structure of different groups within a dataset (Diaz-Papkovich et al., 2021). Additionally, capturing the population structure by inferring admixture proportions using deletions also provided a coherent scenario. Our analysis highlights how, at K = 2, European, Roma and Middle East samples show a major West Eurasian component, while the East/South Asian component increasingly appears moving east, from Middle Eastern samples to South Asians. At K = 3, a Roma component appears as the prevalent ancestry in Roma samples and also in lower proportions in South Asian and European genomes. Our results are a warning for careful filtering when using CNVs and combining different datasets to assess population structure.

Deletions in Roma show a slight relaxation of natural selection

In our analysis, we have observed that the Roma carry more deletions than other European or Asian populations, that this additional load occurs in intergenic and intronic locations (but not in exons), that intergenic deletions in the Roma are longer, and that intronic deletions in the Roma are enriched for genes that are intolerant to loss of function (Lof) mutations. This might be a population or sample-specific artefact in the calling of deletions; however, variables that could affect the calling step, such as genome coverage, do not discriminate exclusively the Roma, as this latter population and the samples from SGDP share similar sequencing depth profiles. Differences in coverage among different batches, indeed, have

been shown to affect CNVs calling in specific regions, but not overall (Khayat et al., 2021). Additionally, these spurious effects are unlikely to result in the bias observed in the Roma towards deletions in introns and intergenic regions.

Although coding variation is the most obvious source of phenotypic differences, the evidence for introns and intergenic regions harbouring functional variation has been accumulating (Keegan et al., 2022; Petersen et al., 2022; Rigau et al., 2019; Telonis & Rigoutsos, 2021; Vaz-Drago et al., 2017). Thus, the additional intronic and intergenic deletions in the Roma point at a slight relaxation of natural selection; the effect of deletions in these regions is likely to be milder than in exons, which, in Roma, do not tolerate deletions at a higher rate than in other populations. The Roma present a unique combination of fragmentation, partial reproductive isolation, but also of admixture with their host populations. Founder events would have accumulated deletions at more tolerated locations with fewer constraints. Admixture, on the other hand, might have introduced new sources of variation in the population, while selection against deleterious mutations still acted to reduce the accumulation of harmful exonic deletions. As shown by reports on worldwide populations, usually selection acts against larger deletions in the genome (Sudmant, Mallick, et al., 2015), however, in our case this result could indicate that less efficient purifying forces may have taken place either because of the population history or because of the intergenic/intronic nature of the variants, bearing a presumably lower disrupting potential. In summary, the putative relaxed purifying selection in closed communities, which has been object of debate and addressed in finer detail by some reports (Balick et al., 2015; W. Fu et al., 2014; Gravel, 2016; Henn et al., 2016), could be detectable only for low-impact mutations, such as intronic ones in the Roma.

Deletions recapitulate the history of the Roma

We show that Roma retain more population-specific variants compared to closely related Europeans and individuals from Middle East, thus exhibiting larger proportions of low-frequency and rare variants. This trend is mostly driven by deletions; indeed, their power to uncover the structure among populations is higher than duplications, whose recurrent mutations over short time spans tend to homogenise variability among individuals, masking past events and making similar variants more prone to be identical by state rather than by descent (Sudmant, Mallick, et al., 2015). The characteristic features of Roma had an influential impact on the apportionment and the accumulation of deleterious variants in the population. Although only few studies addressed underrepresented populations using CNVs to analyse their histories, these reports collectively highlight diversified patterns of allele sharing among populations, within-isolates relatedness, differential admixture with larger populations and signals of selection for CNV loci at genes with putative adaptive phenotypes (W. Chen et al., 2011; R. Fu et al., 2018; H. Lou et al., 2015; Nyangiri et al., 2020). These studies had indeed shown the usefulness and potential of this class of variants in recapitulating a portion of missing genetic dynamics of populations. We argue that, in spite of the population's dynamics, admixture with European and Middle East/Caucasus individuals in a context of an expanding population, over time, may have homogenised the differences between populations and the traces of more remote events, detectable nowadays, might reside at quasi-functional or slightly deleterious variants.

Highly differentiated CNVs in Roma intersect some genes of biomedical interest

Estimating the differentiation for shared CNVs in pairwise population comparisons by means of V_{ST} statistics, we found that intronic variants are significantly more differentiated in pairs with Roma, driving the overall trend. We could identify only one significant frequency difference, between the Roma and Middle East populations, for intronic variants when dividing for intolerant genes categories (LOEUF bin 2), showing Roma as the population with higher frequencies. Nonetheless, we also identified a significant difference in frequencies considering all intolerant categories together (LOEUF 0-4), with Roma exhibiting higher frequencies, even though pairwise populations comparisons did not survive multiple test correction.

We identified five deletions and three duplications (having one singleton variant respectively) overlapping both with known TADs coordinates and genes. Collectively, these CNVs intersect genes with known functions, such as metabolic/physiological processes or immune system, the latter being of particular interest since one deletion found in all populations removes four immunoglobulin genes and a portion of a fifth one. We could not pinpoint any particular population-specific pattern of frequencies for CNVs intersecting TADs; nevertheless, we noticed a slight but consistent increase in frequency, from South Asia to Europe for the deletion overlapping immune-related genes in the immunoglobulin heavy chain gene cluster (*IGHG4*, *IGHG2*, *IGHA1*, *IGHG1* and *IGHG3*), with the Roma exhibiting the highest frequency in our dataset. Previous studies on Roma showed signals of positive selection at loci involved in the immune response related to cytokine production (Dobon et al., 2020; Laayouni et al., 2014). Even though selection signals were not detected in the immunoglobulin heavy chain gene cluster, it is known that CNVs drove the evolution of this locus

and, moreover, deletions might affect immunoglobulin production (Bottarol et al., 1989; Hendriks et al., 1989; Migone et al., 1984; Smith et al., 1989). We only observed a slight increase in the frequency of the deletion in Roma that could fit the hypothesis of a poorer effect of selection or drift at this locus.

Exploring further the possible deleterious nature of our variants, we assessed the levels of LD with known pathogenic SNPs from the GWAS catalog and identified 78 deletions in linkage with 125 trait-associated SNPs. Out of the whole set of these associated diseases, we could highlight four categories including most conditions: metabolic, neurodevelopmental/neurological, developmental and haematological-cardiovascular disorders. Although we acknowledge that only 33 tagged deletions reside on the same gene of the associated SNP (or SNPs), most deletions (41/45) with no common gene are intergenic variants which, among all linked deletions, are those residing in closer proximity to the linked SNP(s) and, thus, might exert a specific influence on the trait-related variant. As an example, the thirty closest deletions sharing no gene with the tagged SNP are all intergenic variants and range in distance from 71bp to 18.8kb. This evidence points at the importance of including intergenic variants in analyses assessing CNV function, as such mutations could be either actors or co-players, modifying their genomic neighbourhood, participating to different scenarios, as already reported for specific diseases (Farrell et al., 2011; Loots et al., 2005; Staehling-hampton et al., 2002; Uyan et al., 2013). Overall, SNPs in LD with intergenic deletions show associations with traits related to development, neurodevelopmental, metabolic and haematological conditions, as well as other traits such as height, smoking behaviour and heart/cardiovascular ones. For genic deletions, it is expected, and probably more likely, that their influence over gene products or regulatory functions would be stronger than intergenic

ones. Together, this set of deletions primarily associate to metabolic/inflammatory, cancer and neurodevelopmental/neurological traits. The collection of conditions related to metabolism mainly pertains to cholesterol levels, type 2 diabetes, alanine transaminase levels and obesity traits. Genes containing SNPs in LD with deletions had low reported LOEUF values, indicating their intolerance to loss of function (*CCDC50*, *JAZF1*, *MYO9A*, *CNOT1* genes having respectively three, one, one and zero LOEUF bin scores). Intriguingly, a previous study showed how European Roma carried higher frequencies of SNPs involved in hyperlipidemia (Mendizabal et al., 2013); we found one deletion in *RHCE* gene in linkage with one cholesterol-associated SNP within the neighbouring *MACO1* gene (however, a direct functional effect upon the *RHCE* gene, which codes for a Rh-like red blood cell antigen, should not be dismissed), and indeed the deletion is higher in frequency within Roma.

Regarding deletions associated with cancers, we specifically found breast, lung and colorectal types, with the latter resulting from rs448513 in LD with an intronic deletion, in *TANC1* gene. This gene shows signs of intolerance to LoF (LOEUF bin: 2) further corroborating the potential disruptive action of intronic variants. Neurodevelopmental/neurological diseases include a vast range of disorders such as schizophrenia, attention deficit hyperactivity disorder (ADHD), bipolar disorder and multiple sclerosis. Once again, two genes, *PCCB* and *VMP1*, hosting deletions in LD with SNPs, show low tolerance to LoF (LOEUF bins: 4 and 1 respectively). With the exception of few examples, generally regarding lower deletions frequencies in the South Asia population linked to metabolic, neurodevelopmental and cancer-associated SNPs, we could not highlight any systematic or marked difference in the frequencies of deletions among the populations. This evidence highlights a general absence of population-specific burden for deletions in linkage with disease-

associated SNPs, or to a diffused appearance of the associated traits among widespread groups. The instances of lower frequencies for deletions in South Asian individuals involve cases of metabolism, cancer-related and neurodevelopmental traits and we could only speculate that different dynamics and population history within Western Eurasians might have favoured higher diffusion of those variants.

Genes intersected by CNVs in Roma are enriched for central nervous system functions

We discovered that Roma carry a marked prevalence of GO terms associated to common functions subsets of inputted genes lists. Intriguingly, we could highlight marked differences only when using this type of gene sets, i.e., intolerant genes that contained intronic deletions, and not while using other sets, such as private deletions within populations or general classification based on genomic location. This is unlikely the result of a general higher number of deletions in Roma but rather the specific function of the affected genes. Roma show more biological process GO terms in each defined category (Nervous system, Signalling and Development categories plus "Other" containing general unrelated terms) compared to the other populations, a strong difference can be noticed for the Nervous system and Signalling categories. We find these results of particular interest in light of the known private diseases specifically affecting Roma people in Europe. Indeed, among the different types of private disease-causing mutations described in the Roma, some involve neuropathies and neurological diseases such as hereditary motor and sensory neuropathy-Lom/Russe types, congenital cataracts facial dysmorphism neuropathy and limb-girdle muscular dystrophy type 2C (Angelicheva et al., 1999; Kalaydjieva et al., 1996, 2001, 2005; Morar et al., 2004). Nervous system-related GO terms often involved neurons connections organization, synaptic communication or brain development,

highlighting the presence of putatively deleterious variants affecting physiological neuronal functions particularly in Roma, in line with previous reports of a higher rate of slightly deleterious variants, for other disorders, in Roma individuals (Mendizabal et al., 2013). Moreover, the disease-associated SNPs assessed in (Mendizabal et al., 2013) reside in genes belonging to biological processes associated to the significant GO terms we identified in our analysis, highlighting a possible action of different markers (deletions and SNPs) within same sets of genes, specifically affecting their functions.

Isolated populations are an under analysed genomic resource, also for CNVs

Populations of non-European descent have traditionally been understudied in the context of genetic variation, particularly favouring GWAS research on more accessible cohorts of general European ancestry (Bustamante et al., 2011; Popejoy & Fullerton, 2016). Ironically, what should be one important goal of human genetics research: uncovering an increasingly clearer and more complete picture of human genetic variation worldwide, portray a fairer representation of different human populations and advancing current knowledge on genetic diseases using diverse sets of populations (Zeggini, 2014), has often been disregarded in favour of a Eurocentric perspective (Need & Goldstein, 2009; Sirugo et al., 2019). Numerous studies addressing population isolates, indeed, contributed significantly to identify the loci underlying complex diseases: bipolar disorder and schizophrenia in Finland and Basque populations (Palo et al., 2007; Parsons et al., 2007), studies on Iceland individuals highlighting variants associated to atrial fibrillation, myocardial infarction, type 2 diabetes and glaucoma (Gudbjartsson et al., 2007; Helgadottir et al., 2007; Manolescu et al., 2004; Steinthorsdottir et al., 2007; Thorleifsson et al., 2007) and also traits as height and pigmentation in Finland, Iceland,

Sardinia and Amish populations (Gudbjartsson et al., 2008; Sanna et al., 2008; Sulem et al., 2007, 2008). It has been suggested that addressing isolated populations for studying diseases can help in reducing the variance of environmental variables on pathogenic conditions, as homogeneity in phenotype and environment within isolates would facilitate the disease-gene recognition (Kristiansson et al., 2008), thus favouring the inclusion of underrepresented populations to advance our understating of health-related traits.

Data availability statement

Upon acceptance of this manuscript, a vcf file containing the CNV calls for the samples analysed will be deposited in the European Genome-Phenome Archive (EGA).

Acknowledgments

This work was supported by the Spanish Ministry of Economy and Competitiveness and *Agencia Estatal de Investigación* (grant numbers CGL2016-75389-P (MINEICO/FEDER, UE), PID2019-106485GB-I00/AEI/10.13039/501100011033 (MINEICO), and "Unidad María de Maeztu"(CEX2018-000792-M) to FC and DC; and Agència de Gestió d'Ajuts Universitaris i de la Recerca (Generalitat de Catalunya, grant 2017SGR00702).

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21:974–984.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655–1664.

Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, Chen Y, Hurles ME, Tyler-Smith C, Xue Y. 2020. Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell* 182:189-199.e15.

Angelicheva D, Turnev I, Dye D, Chandler D, Thomas PK, Kalaydjieva L. 1999. Congenital cataracts facial dysmorphism neuropathy (CCFDN) syndrome: A novel developmental disorder in Gypsies maps to 18qter. *European Journal of Human Genetics* 7:560–566.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: Tool for the unification of biology. *Nat Genet* 25:25–29.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AME, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* [Internet] 176:663-675.e19. Available from: http://dx.doi.org/10.1016/j.cell.2018.12.019

Ayub Q, Tyler-Smith C. 2009. Genetic variation in South Asia: Assessing the influences of geography, language and ethnicity for understanding history and disease risk. *Brief Funct Genomic Proteomic* 8:395–404.

Balick DJ, Do R, Cassa CA, Reich D, Sunyaev SR. 2015. Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck. *PLoS Genet* 11:1–23.

Beagan JA, Phillips-Cremins JE. 2020. On the existence and functionality of topologically associating domains. *Nat Genet* 52:8–16.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. Pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* 32:2817–2823.

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science (1979)* 367.

Bianco E, Laval G, Font-Porterias N, García-Fernández C, Dobon B, Sabido-Vera R, Sukarova Stefanovska E, Kučinskas V, Makukh H, Pamjav H, et al. 2020. Recent common origin, reduced population size, and marked admixture have shaped European roma genomes. *Mol Biol Evol* 37:3175–3187.

Boerger BH. 1984. Proto-Romanes Phonology. 8421721.

Bottarol A, De Marchi M, De Langea G, Boccazzi C, Caldesiz F, Gallinat R, Carbonara AO. 1989. New types of multiple and single gene deletions in the human IgCH locus. *Immunogenetics* 29:44–48.

Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 47:D1005–D1012.

Bustamante CD, De La Vega FM, Burchard EG. 2011. Genomics for the world. *Nature* 475:163–165.

Cabrera-Serrano M, Mavillard F, Biancalana V, Rivas E, Morar B, Hernández-Laín A, Olive M, Muelas N, Khan E, Carvajal A, et al. 2018. A Roma founder BIN1 mutation causes a novel phenotype of centronuclear myopathy with rigid spine. *Neurology* 91:e339–e348.

Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, Grenier JC, Gbeha E, Hamdan FF, Girard S, et al. 2013. Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. *PLoS Genet* 9.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 6:677–681.

Chen W, Hayward C, Wright AF, Hicks AA, Vitart V, Knott S, Wild SH, Pramstaller PP, Wilson JF, Rudan I, et al. 2011. Copy number variation across European populations. *PLoS One* 6.

De Cid R, Riveira-Munoz E, Zeeuwen PLJM, Robarge J, Liao W, Dannhauser EN, Giardina E, Stuart PE, Nair R, Helms C, et al. 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet* 41:211–215.

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera A v., Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* 581:444–451.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* [Internet] 464:704–712. Available from: http://dx.doi.org/10.1038/nature08516

Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, et al. 2022. Ensembl 2022. *Nucleic Acids Res* 50:D988–D995.

Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. *Nat Commun* 10.

Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, Antonacci F, Penewit K, Denman L, Raja A, et al. 2017. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol* 1:1–10.

Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, Yu K, Rubanova Y, Macintyre G, Demeulemeester J, et al. 2021. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* 184:2239-2254.e39.

Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. 2021. A review of UMAP in population genetics. *J Hum Genet* 66:85–91.

Dobon B, ter Horst R, Laayouni H, Mondal M, Bianco E, Comas D, Ioana M, Bosch E, Bertranpetit J, Netea MG. 2020. The shaping of immunological responses through natural selection after the Roma Diaspora. *Sci Rep* 10:1–12.

Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, Gudbjartsson DF, Stefansson K, Halldorsson B V., Melsted P. 2019. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun* 10:1–8.

Farrell JJ, Sherva RM, Chen Z, Luo H, Chu BF, Ha SY, Li Chi Kong, Lee ACW, Li RCH, Li Chi Keung, et al. 2011. A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. 117:4935–4945.

Firth H V., Wright CF. 2011. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* 53:702–703.

Font-Porterias N, Arauna LR, Poveda A, Bianco E, Rebato E, Prata MJ, Calafell F, Comas D. 2019. European Roma groups show complex West Eurasian admixture footprints and a common South Asian genetic origin. *PLoS Genet* 15:e1008417.

Font-Porterias N, Giménez A, Carballo-Mesa A, Calafell F, Comas D. 2021. Admixture Has Shaped Romani Genetic Diversity in Clinically Relevant Variants. *Front Genet* 12:1–12.

Fox, John & Weisberg S. 2011. An R Companion to Applied Regression.

Fraser A. 1992. The gypsies. Oxford: Wiley-Blackwell

Fu R, Mokhtar SS, Phipps ME, Hoh BP, Xu S. 2018. A genome-wide characterization of copy number variations in native populations of Peninsular Malaysia. *European Journal of Human Genetics* 26:886–897.

Fu W, Gittelman RM, Bamshad MJ, Akey JM. 2014. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am J Hum Genet* 95:421–436.

García-Fernández C, Font-Porterias N, Kučinskas V, Sukarova-Stefanovska E, Pamjav H, Makukh H, Dobon B, Bertranpetit J, Netea MG, Calafell F, et al. 2020. Sex-biased patterns shaped the genetic history of Roma. *Sci Rep* 10.

Gautam P, Jha P, Kumar D, Tyagi S, Varma B, Dash D, Mukhopadhyay A, Mukerji M. 2012. Spectrum of large copy number variations in 26 diverse Indian populations: Potential involvement in phenotypic diversity. *Hum Genet* 131:131–143.

Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, Boerwinkle E, Gibbs RA, Sing CF, Clark AG, et al. 2014. Neutral genomic regions

refine models of recent rapidhuman population growth. *Proc Natl Acad Sci U S A* 111:757–762.

Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. 2016. RegioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* 32:289–291.

Geoffroy V, Guignard T, Kress A, Gaillard JB, Solli-Nowlan T, Schalk A, Gatinois V, Dollfus H, Scheidecker S, Muller J. 2021. AnnotSV and knotAnnotSV: A web server for human structural variations annotations, ranking and analysis. *Nucleic Acids Res* 49:W21–W28.

Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, Muller J. 2018. AnnotSV: An integrated tool for structural variations annotation. *Bioinformatics* 34:3572–3574.

Girirajan S, Dennis MY, Baker C, Malig M, Coe BP, Campbell CD, Mark K, Vu TH, Alkan C, Cheng Z, et al. 2013. Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am J Hum Genet* 92:221–237.

Graffelman J. 2015. Exploring diallelic genetic markers: The HardyWeinberg package. *J Stat Softw* 64:1–23.

Granot Y, Tal O, Rosset S, Skorecki K. 2016. On the apportionment of population structure. *PLoS One* 11.

Gravel S. 2016. When is selection effective? *Genetics* 203:451–462.

Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, Angelicheva D, Calafell F, Oefner PJ, Shen P, et al. 2001. Origins and divergence of the Roma (gypsies). *Am J Hum Genet* 69:1314–1331.

Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, Jonasdottir A, Baker A, Thorleifsson G, Kristjansson K, et al. 2007. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 448:353–357.

Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson B V., Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, et al. 2008. Many sequence variants affecting diversity of adult human height. *Nat Genet* 40:609–615.

Hamdan A, Ewing A. 2022. Unravelling the tumour genome: The evolutionary and clinical impacts of structural variants in tumourigenesis. *Journal of Pathology* 257:479–493.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33.

Handsaker RE, van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, Mccarroll SA. 2015. Large multiallelic copy number variations in humans. *Nat Genet* 47:296–303.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* 43:269–276.

Hao W, Storey JD. 2019. Extending tests of hardy-weinberg equilibrium to structured populations. *Genetics* 213:759–770.

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* 7:1–10.

Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir Aslaug, Jonasdottir Abalbjorg, Sigurdsson A, Baker A, Palsson A, et al. 2007. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Obstet Gynecol Surv* 62:585–587.

Hendriks RW, Van Tol MJD, De Lange GG, Schuurman RKB. 1989. Inheritance of a Large Deletion within the Human Immunoglobulin Heavy Chain Constant Region Gene Complex and Immunological Implications. *Scand J Immunol* 29:535–541.

Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, et al. 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci U S A* 113:E440–E449.

Hollox EJ, Zuccherato LW, Tucci S. 2021. Genome structural variation in human evolution. *Trends in Genetics* [Internet]:1–14. Available from: https://doi.org/10.1016/j.tig.2021.06.015

Hurles ME, Dermitzakis ET, Tyler-Smith C. 2008. The functional impact of structural variation in humans. *Trends in Genetics* 24:238–245.

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, et al. 2008. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84:148–161.

Kaklamani VG, Wisinski KB, Sadim M, Gulden C, Do A, Offit K, Baron JA, Ahsan H, Mantzoros C, Pasche B. 2008. Variants of the adiponectin (ADIPOQ) and adiponectin receptor 1 (ADIPOR1) genes and colorectal cancer risk. *JAMA - Journal of the American Medical Association* 300:1523–1531.

Kalaydjieva L, Gresham D, Calafell F. 2001. Genetic studies of the Roma (Gypsies): A review. *BMC Med Genet* 2.

Kalaydjieva L, Hallmayer J, Chandler D, Savov A, Nikolova A, Angelicheva D, King RHH, Ishpekova B, Honeyman K, Calafell F, et al. 1996. Gene mapping in Gypsies identifies a novel demyelinating neuropathy on chromosome 8q24. *Nat Genet* 14:214–217.

Kalaydjieva L, Morar B, Chaix R, Tang H. 2005. A newly discovered founder population: The Roma/Gypsies. *BioEssays* 27:1084–1094.

Kalaydjieva L, Perez-Lezaun A, Angelicheva D, Onengut S, Dye D, Bosshard NU, Jordanova A, Savov A, Yanakiev P, Kremensky I, et al. 1999. A founder mutation in the GK1 gene is responsible for galactokinase deficiency in Roma (Gypsies). *Am J Hum Genet* 65:1299–1307.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581:434–443.

Kato H, Kimura H, Kushima I, Takahashi N, Aleksic B, Ozaki N. 2022. The genetic architecture of schizophrenia: review of large-scale genetic studies. *J Hum Genet*.

Keegan NP, Wilton SD, Fletcher S. 2022. Analysis of Pathogenic Pseudoexons Reveals Novel Mechanisms Driving Cryptic Splicing. *Front Genet* 12.

Khayat MM, Mohammad S, Sahraeian E, Zarate S, Carroll A, Hong H, Pan B, Shi L, Gibbs RA, Mohiyuddin M, et al. 2021. Hidden biases in germline structural variant detection. *Genome Biol* 22:347.

Kimura M, Maruiama T, Crow JF. 1963. The Mutation Load in Small Populations. *Genetics* 48:1303–1312.

Kimura M, Ohta T. 1969. The Average Number of Generations Until Fixation of a Mutant Gene in a Finite Population. *Genetics* 61:763–771.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* 20:8–11.

Kristiansson K, Naukkarinen J, Peltonen L. 2008. Isolated populations and complex disease gene identification. *Genome Biol* 9.

Laayouni H, Oosting M, Luisi P, Ioana M, Alonso S, Ricano-Ponce I, Trynka G, Zhernakova A, Plantinga TS, Cheng SC, et al. 2014. Convergent evolution in European and Rroma populations reveals pressure exerted by plague on Toll-like receptors. *Proc Natl Acad Sci U S A* 111:2668–2673.

Laso-Jadart R, Harmant C, Quach H, Zidane N, Tyler-Smith C, Mehdi Q, Ayub Q, Quintana-Murci L, Patin E. 2017. The Genetic Legacy of the Indian Ocean Slave Trade: Recent Admixture and Post-admixture Selection in the Makranis of Pakistan. *Am J Hum Genet* 101:977–984.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* 9.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol* 15:1–19.

Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019a. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* 47:W199–W205.

Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019b. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* 47:W199–W205.

Liégeois J-P. 1994. Roma, Gypsies, Travellers.

Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, Esko T, Mägi R, Inouye M, Lappalainen T, et al. 2014. Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genet* 10.

Linck E, Battey CJ. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol Ecol Resour* 19:639–647.

Lohmueller KE. 2014. The distribution of deleterious genetic variation in human populations. *Curr Opin Genet Dev* 29:139–146.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451:994–997.

Loots GG, Kneissel M, Keller H, Baptist M, Chang J, Collette NM, Ovcharenko D, Plajzer-Frick I, Rubin EM. 2005. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res* 15:928–935.

Lou H, Li S, Jin W, Fu R, Lu D, Pan X, Zhou H, Ping Y, Jin L, Xu S. 2015. Copy number variations and genetic admixtures in three Xinjiang ethnic minority groups. *European Journal of Human Genetics* 23:536–542.

Lupiáñez DG, Spielmann M, Mundlos S. 2016. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics* 32:225–237.

Lutz BD. 1995. Gypsies as victims of the holocaust. *Holocaust Genocide Stud* 9:346–359.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538:201–206.

Manolescu A, Helgadottir A, Kong A, Valdimarsson EM, Topol EJ, Thorleifsson G, Gudmundsson G, Thorgeirsson G, Hakonarson H, Johannsson H, et al. 2004. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat Genet* 36:233–239.

McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, et al. 2008. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* 40:1107–1112.

McInnes L, Healy J, Melville J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Melegh BI, Banfai Z, Hadzsiev K, Miseta A, Melegh B. 2017. Refining the South Asian Origin of the Romani people. *BMC Genet* 18:1–13.

Mendizabal I, Lao O, Marigorta UM, Kayser M, Comas D. 2013. Implications of Population History of European Romani on Genetic Susceptibility to Disease. *Hum Hered* 76:194–200.

Migone N, Oliviero S, de Lange G, Delacroix DL, Boschis D, Altruda F, Silengo L, DeMarchi M, Carbonara AO. 1984. Multiple gene deletions within the human immunoglobulin heavy-chain cluster. *Proc Natl Acad Sci U S A* 81:5811–5815.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65.

Milton S. 1991. Gypsies and the Holocaust. *Hist Teacher* 24:375.

Mohamad Isa II, Jamaluddin J, Achim NH, Abubakar S. 2020. Population-specific profiling of CCL3L1 copy number of the three major ethnic groups in Malaysia and the implication on HIV susceptibility. *Gene* 754:144821.

Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, Comas D, Laayouni H, Li Q, Majumder PP, et al. 2016. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat Genet* 48:1066–1070.

Moorjani P, Patterson N, Loh PR, Lipson M, Kisfali P, Melegh BI, Bonin M, Kádaši Ľ, Rieß O, Berger B, et al. 2013. Reconstructing Roma History from Genome-Wide Data. *PLoS One* 8.

Morar B, Gresham D, Angelicheva D, Tournev I, Gooding R, Guergueltcheva V, Schmidt C, Abicht A, Lochmuller H, Tordai A, et al. 2004. Mutation history of the roma/gypsies. *Am J Hum Genet* 75:596–609.

Moreno-Cabrera JM, del Valle J, Castellanos E, Feliubadaló L, Pineda M, Serra E, Capellá G, Lázaro C, Gel B. 2021. CNVfilteR: an R/Bioconductor package to identify false positives produced by germline NGS CNV detection tools. *Bioinformatics*:1–3.

Morris-Rosendahl DJ, Crocq M-A. 2020. Neurodevelopmental disorders—the history and future of a diagnostic concept . *Dialogues Clin Neurosci* 22:65–72.

Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, Bhavani GS, Girisha KM, Mustak MS, Srinivasan S, et al. 2017. The promise of discovering population-specific disease-associated genes in South Asia. *Nat Genet* 49:1403–1407.

Narang A, Jha P, Kumar D, Kutum R, Mondal AK, Dash D, Mukerji M. 2014. Extensive copy number variations in admixed Indian population of African ancestry: Potential involvement in adaptation. *Genome Biol Evol* 6:3171–3181.

Need AC, Goldstein DB. 2009. Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics* 25:489–494.

Nyangiri OA, Noyes H, Mulindwa J, Ilboudo H, Kabore JW, Ahouty B, Koffi M, Asina OF, Mumba D, Ofon E, et al. 2020. Copy number variation in human genomes from three major ethno-linguistic groups in Africa. *BMC Genomics* 21:1–15.

Palo OM, Antila M, Silander K, Hennah W, Kilpinen H, Soronen P, Tuulio-Henriksson A, Kieseppä T, Partonen T, Lönnqvist J, et al. 2007. Association of distinct allelic haplotypes of DISC1 with psychotic and bipolar spectrum disorders and with underlying cognitive impairments. *Hum Mol Genet* 16:2517–2528.

Parsons MJ, Mata I, Beperet M, Iribarren-Iriso F, Arroyo B, Sainz R, Arranz MJ, Kerwin R. 2007. A dopamine D2 receptor gene-related polymorphism is associated with schizophrenia in a Spanish population isolate. *Psychiatr Genet* 17:159–163.

Patterson MD, Marschall T, Pisanti N, Van Iersel L, Stougie L, Klau GW, Schönhuth A. 2015. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of Computational Biology* 22:498–509.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* 2:2074–2093.

Pedersen CET, Lohmueller KE, Grarup N, Bjerregaard P, Hansen T, Siegismund HR, Moltke I, Albrechtsen A. 2017. The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: Insights from the Greenlandic Inuit. *Genetics* 205:787–801.

Pembleton LW, Cogan NOI, Forster JW. 2013. StAMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Mol Ecol Resour* 13:946–952.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, et al. 2008. Copy number variation and evolution in humans and chimpanzees. *Genome Res* 18:1698–1710.

Petersen USS, Doktor TK, Andresen BS. 2022. Pseudoexon activation in disease by non-splice site deep intronic sequence variation — wild type pseudoexons constitute high-risk sites in the human genome. *Hum Mutat* 43:103–127.

Piccolo F, Jeanpierre M, Leturcq F, Dodé C, Azibi K, Toutain A, Merlini L, Jarre L, Navarro C, Krishnamoorthy R, et al. 1996. A founder mutation in the γ-sarcoglycan gene of Gypsies possibly predating their migration out of India. *Hum Mol Genet* 5:2019–2022.

Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature* 538:161–164.

R Development Core Team. 2003. R: A language and environment for statistical computing.

Rainer J. 2017. EnsDb.Hsapiens.v79: Ensembl based annotation package. R package version 2.99.0. *R package* version 2.:1–6.

Rainer J, Gatto L, Weichenberger CX. 2019. Ensembldb: An R package to create and use Ensembl-based annotation resources. *Bioinformatics* 35:3151–3153.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* 444:444–454.

Reyniers A. 1995. Gypsy populations and their movements within Central and Eastern Europe and towards some OECD countries.

Rigau M, Juan D, Valencia A, Rico D. 2019. Intronic CNVs and gene expression variation in human populations. *PLoS Genet* 15:1–23.

Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, Shen H, Timpson N, Lettre G, Usala G, et al. 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* 40:198–203.

Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA. 2011. A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostatistics* 12:33–50.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. *Science (1979)* 316:445–449.

Sekiguchi M, Sobue A, Kushima I, Wang C, Arioka Y, Kato H, Kodama A, Kubo H, Ito N, Sawahata M, et al. 2020. ARHGAP10, which encodes Rho GTPase-activating protein 10, is a novel gene for schizophrenia risk. *Transl Psychiatry* 10:247.

Serres-Armero A, Davis BW, Povolotskaya IS, Morcillo-Suarez C, Plassais J, Juan D, Ostrander EA, Marques-Bonet T. 2021. Copy number variation underlies complex phenotypes in domestic dog breeds and other canids. *Genome Res* 31:762–774.

Shan MA, Meyer OS, Refn M, Morling N, Andersen JD, Børsting C. 2021. Analysis of skin pigmentation and genetic ancestry in three subpopulations from pakistan: Punjabi, pashtun, and baloch. *Genes (Basel)* 12.

Shanta O, Noor A, Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, et al. 2020. The effects of common structural variants on 3D chromatin structure. *BMC Genomics* 21:1–10.

Singh T, Walters JTR, Johnstone M, Curtis D, Suvisaari J, Torniainen M, Rees E, Iyegbe C, Blackwood D, McIntosh AM, et al. 2017. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat Genet* 49:1167–1173.

Sirugo G, Williams SM, Tishkoff SA. 2019. The Missing Diversity in Human Genetic Studies. *Cell* 177:26–31.

Smith CIE, Hammarstrom L, Henter J-I, De Lange GG. 1989. Molecular and serologic analysis of IgG1 deficiency caused by new forms of the constant region of the Ig H chain gene deletions. *Journal of Immunology* 142.

Soylev A, Kockan C, Hormozdiari F, Alkan C. 2017. Toolkit for automated and rapid discovery of structural variants. *Methods* 129:3–7.

Sridhar CR. 2006. Historical Amnesia: The Romani Holocaust. *Econ Polit Wkly* [Internet] 41:3569–3571. Available from: http://www.jstor.org.proxy.library.georgetown.edu/stable/4418585

Staehling-hampton K, Proll S, Paeper BW, Zhao L, Charmley P, Brown A, Gardner JC, Galas D, Schatzman RC, Beighton P, et al. 2002. A 52-kb Deletion in the SOST-MEOX1 Intergenic Region on 17q12-q21 Is Associated With van Buchem Disease in the Dutch Population. 152:144–152.

Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* 455:232–236.

Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, et al. 2007. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 39:770–775.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazlsy C, Thorne N, Redon R, Bird CP, De Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene phenotypes. *Science (1979)* 315:848–853.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science (1979)* 349.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81.

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G, et al. 2008. Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet* 40:835–837.

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, et al. 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* 39:1443–1452.

Telonis AG, Rigoutsos I. 2021. The transcriptional trajectories of pluripotency and differentiation comprise genes with antithetical architecture and repetitive-element content. *BMC Biol* 19:1–19.

Thaler A, Ash E, Gan-Or Z, Orr-Urtreger A, Giladi N. 2009. The LRRK2 G2019S mutation as the cause of Parkinson's disease in Ashkenazi Jews. *J Neural Transm* 116:1473–1482.

Thorleifsson G, Magnusson KP, Sulem P, Walters GB, Gudbjartsson DF, Stefansson H, Jonsson T, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Stefansdottir G, et al. 2007. Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science (1979)* 317:1397–1400.

Urnikyte A, Domarkiene I, Stoma S, Ambrozaityte L, Uktveryte I, Meskiene R, Kasiulevičius V, Burokiene N, Kučinskas V. 2016. CNV analysis in the Lithuanian population. *BMC Genet* 17:1–8.

Uyan Ö, Ömür Ö, Ağim ZS, Özoğuz A, Li H, Parman Y, Deymeer F, Oflazer P, Koç F, Tan E, et al. 2013. Genome-Wide Copy Number Variation in Sporadic Amyotrophic Lateral Sclerosis in the Turkish Population: Deletion of EPHA3 Is a Possible Protective Factor. *PLoS One* 8.

Valls-Margarit J, Galván-Femenía I, Matías-Sánchez D, Blay N, Puiggròs M, Carreras A, Salvoro C, Cortés B, Amela R, Farre X, et al. 2022. GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *Nucleic Acids Res*:1–16.

Vandepoele K, Van Roy N, Staes K, Speleman F, Van Roy F. 2005. A novel gene family NBPF: Intricate structure generated by gene duplications during primate evolution. *Mol Biol Evol* 22:2265–2274.

Vaz-Drago R, Custódio N, Carmo-Fonseca M. 2017. Deep intronic mutations and human disease. *Hum Genet* 136:1093–1111.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat Rev Genet* 14:125–138.

Willer CJ, Speliotes EK, Loos RJF, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C, et al. 2009. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 41:25–34.

Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, Zhou Q, Pan F, Chen Y, Zhang ZX, Dong SS, et al. 2008. Genome-wide Copy-Number-Variation Study Identified a Susceptibility Gene, UGT2B17, for Osteoporosis. *Am J Hum Genet* 83:663–674.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25:2865–2871.

Zeggini E. 2014. Using genetically isolated populations to understand the genomic basis of disease. *Genome Med* 6:12–14.

Zhang B, Kirov S, Snoddy J. 2005. WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* 33.

Tables and figures

	Average counts per genome				
Software	**Total CNVs**	**DEL**	**DUP**	**INS**	**INV**
CNVnator	1753	1356	397	-	-
BreakDancer	2784	2717	-	62	5
Pindel	1184	1059	114	11	-
Tardis	822	650	114	7	40
Lumpy	2290	1974	243	-	73
GenomeSTRiP	1893	1525	368	-	-

Table 1. Average CNVs called per genome for each software. DEL: deletions, DUP: duplications, INS: insertions, INV: inversions

Population	**N common**	**N low frequency**	**N rare**
Roma	1967	479	288
Europe	1899	345	223
Middle East	1835	289	230
South Asia	2006	531	382

Table 2. Distribution of CNVs for frequency class among populations

CNV type	**Exonic**	**Intronic**	**Intergenic**	**Total**
Deletions	211 (6.7%)	1111 (35.0%)	1849 (58.3%)	3171
Duplications	89 (18.2%)	134 (27.4%)	266 (54.4%)	489
Total	300	1245	2115	3660

Table 3. Number of identified deletions and duplications per genomic location. Percentages are over type of CNV

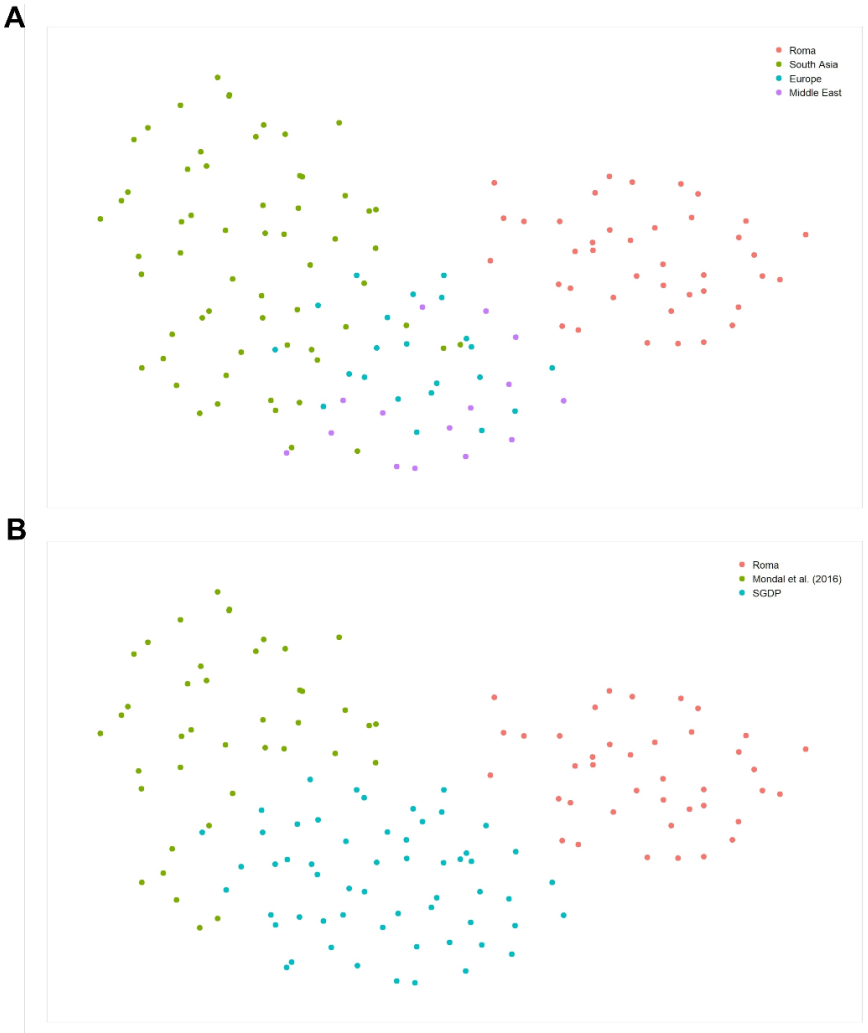


Figure 12. UMAP projection for samples. UMAP plots for deletions copy numbers. UMAP plots representing samples dataset labelled with regional assignation (A) and dataset of origin (B).

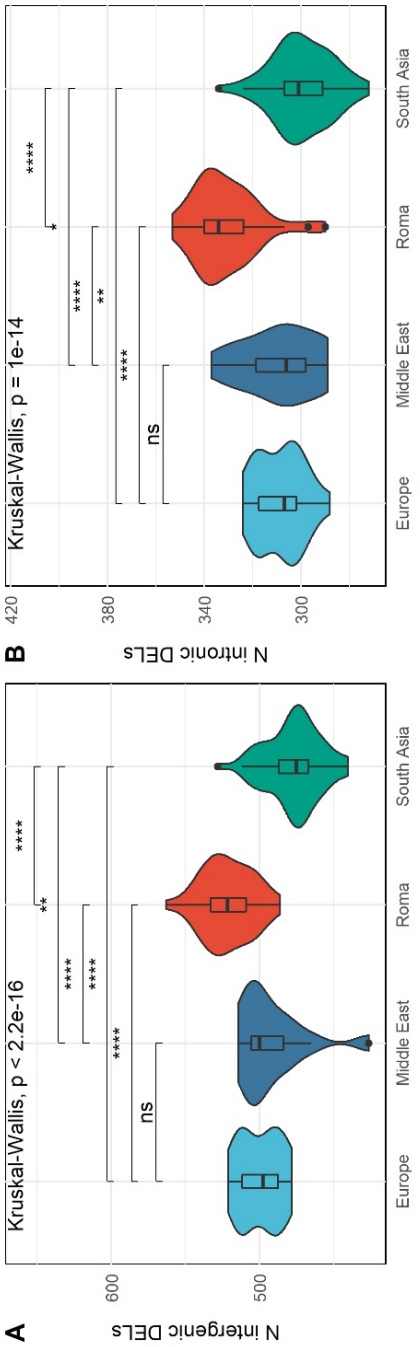


Figure 13. Abundance distribution and statistical tests results for deletions among populations. Statistical test and multiple comparisons results for intergenic (A) and intronic (B) deletions and their relative number distribution among populations.

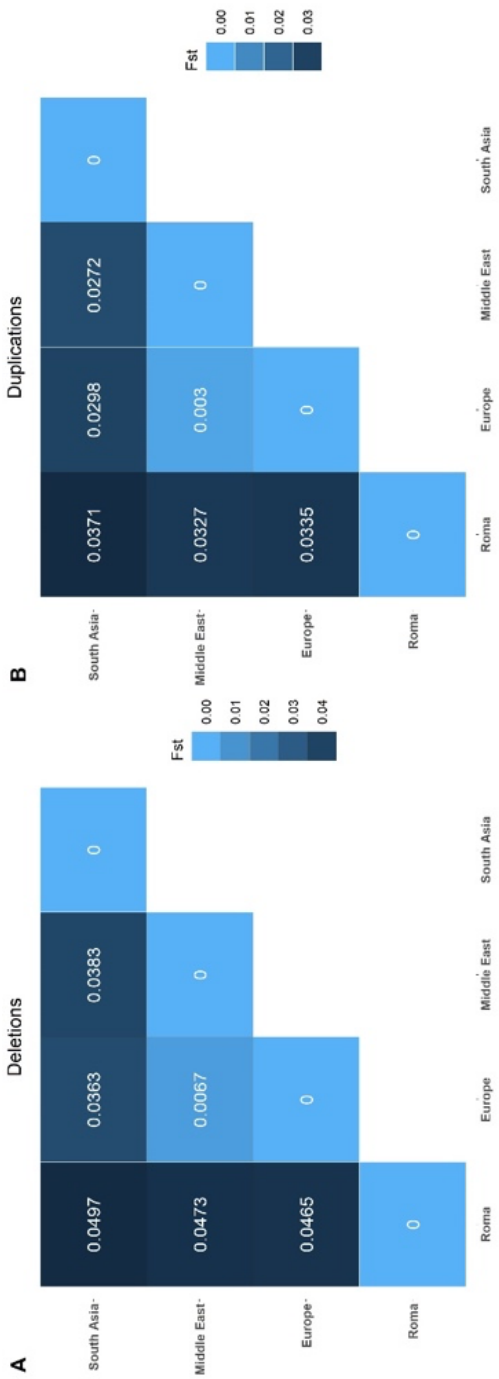


Figure 14. Fst values for pairs of populations. For each pair of population, genome-wide Fst values are shown for deletions (A) and duplications (B)

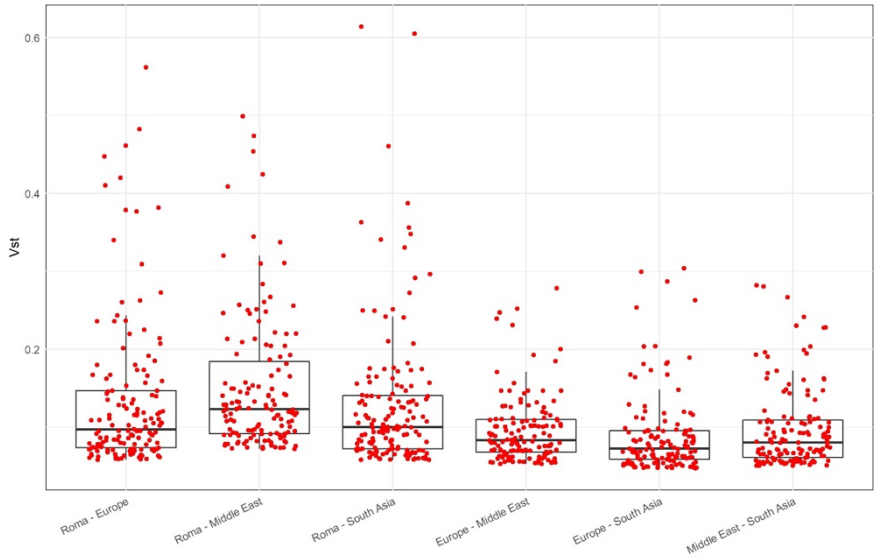


Figure 15. Top quintile Vst values distribution for deletions. Plot showing the distribution of shared deletion in each pair of population. Only intronic variants are displayed

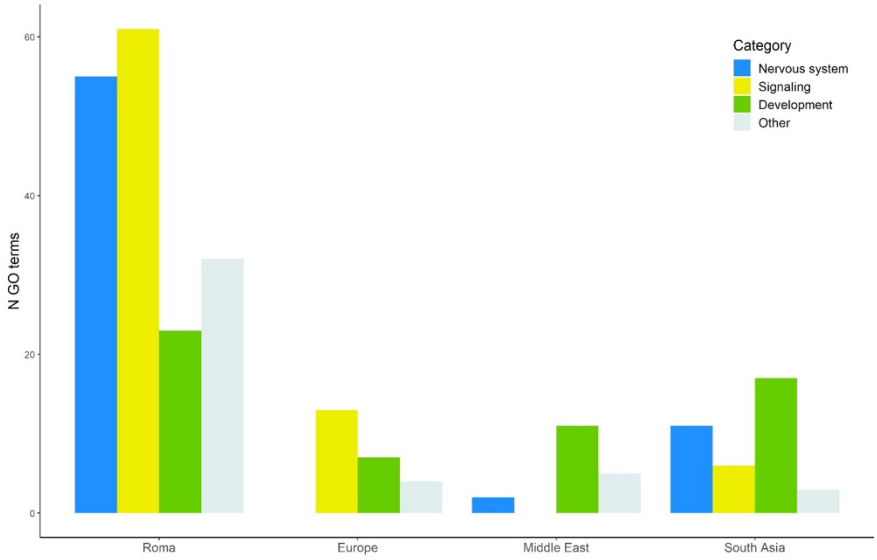


Figure 16. Number of categorized GO terms among populations.

4. DISCUSSION

4.1 What lies before the analyses: the construction of a dataset

Within the context of structural variation analysis in human populations, we highlighted throughout this thesis how the evolution of technology and of the methodologies to better exploit this resource had a strong influence upon the results spanning almost two decades of research. Even though population-scale studies focussing on whole genome analysis of SVs cannot be performed with wet-lab techniques, and indeed are carried out using WGS approaches (greatly gaining from both traditional and newer techniques), they still have to pay the price of the increased levels of complexity resulting from these methods. WGS is the outcome of a high throughput sequencing method or, to give a more figurative indication, massive parallel sequencing and the (vast) amount of biological data generated by genome fragmentation and amplification, packed as paired-end reads information, needs to be digitally handled. While SNP calling methods are quite standardized, SV detection relies on different methodologies implemented by tens of different pieces of software which are not guaranteed to provide the same result. In fact, during the early stages of this project, while progressively framing a clearer picture of the state of the art, it became clear (particularly based on bibliographical research) that the most advocated solution was to leverage different software in an effort to counterbalance the different detecting accuracies of each algorithm, while retaining redundant shared (a proxy for more likely) information. The precise characteristics of the methodologies utilized by different algorithms have been described in the Introduction (see 1.1.3 Detecting structural variants), thus they are not going to be covered extensively here.

To start obtaining the foundation of our dataset, the raw CNV calls, we selected those algorithms that had been described to provide solid information (at least for their stand-alone capabilities) and also that had

been used in projects concerning population genetics of CNVs. The software selection itself was not trivial, as candidate algorithms, scrutinized through published studies, had to be installed on our cluster of computer nodes (as the majority of these software were new and not used by other research groups), tested with one or few samples to assess their performances and, only then, used for all the included samples. Nonetheless, the aforementioned scenario exemplifies an ideal situation of a fixed pipeline, where going from A to B and finally C is straightforward and consequential. The most likely scenario, though, involved solving problems in virtually every step of the process, sometimes related to learning a software itself but more often involving issues with the functioning thereof, leading to contacting the authors in occasionally long epistolary exchanges to address the problems. Fixing errors might require the conversion of input formats (sometimes raw sequencing or mapped files, tens to hundreds of Gb in size for each sample) or formatting the input in different ways, re-run the analyses and assess the results. All these steps intrinsically required resources, in terms of computational work, scripting and processing times, further influencing the generation of the dataset. An additional factor influencing the set of software to use was the constant update rate, either within versions of the same algorithm implementing new features or newly released software (often advertised as the most promising ones) that kept being announced for public usage, thus potentially generating a spiral of testing and implementation to keep adding newer alluring tools to the set. In the end, we decided to stick with a six-software set composed of single method and multi-methods programmes, thus implementing the strength of specialized information extraction (some software also worked in combination with the results from the others) and the broader spectrum of calling methods from multi-algorithm approaches. If a set of six software sounds like adding too much complexity to the process, as stated before in the thesis, studies conducted to date adopted

anywhere between 3 to ~20 calling software for their projects (Ho et al. 2019; Kosugi et al. 2019) precisely to leverage the different characteristics of such a diverse array of detection properties with respect to SNP pipelines.

Once the established set of software was used, attention had to be given to the way each caller inferred the genotypes of the detected variants; indeed, depending on the methodology adopted, similarly to the coordinate concordance, also genotypes are not expected to always match among software. Some of those, by leveraging read-depth information provided direct integer copy number estimation, while others yielded the canonical binary classification[20]. Some callers inferred genotypes as built-in step during the main software operations, while others relied on additional steps provided by companion genotyping software. Additionally, accurate detection of genotypes varies among software, and some could be more confidently trusted than others. Eventually, as it will be discussed below, due to subsequent filtering and resolution of problems, a general regenotyping step had to be included for all the calls, thus suggesting to accurately separating the detection phase to the genotyping one.

After completing the calling steps for each software within each analysed sample, data had to be merged together to create a consensus callset of CNVs in all individuals. Unfortunately, the differences between SNP and SV calling pipelines, also apply to merging procedures. First, SNPs can benefit from online resources providing univocal rsID numbers that can greatly help when joining, subsetting or filtering SNPs sets, secondly, each

[20] The read-depth-based method is implemented in CNVnator calling deletions and duplications. Its algorithm normalizes the average genome coverage, matching the depletions or increments of mapping reads at different loci as lower or higher copy numbers departing from the baseline level 2.
Other callers, leveraging the locations and the way reads map to a putative CNV locus, provide canonical genotype assignations. As for SNPs in classical vcf files, individuals homozygous for the reference allele are marked as “0/0”, heterozygous samples as “0/1” and homozygous for the alternative allele as “1/1”.

SNP has a univocal position in the genome that is used by different software to conveniently merge multiple samples together looking at this information. Again, SVs bear an intrinsic additional level of complexity, being composed of a start and an end position, which hardly match at the base-pair level among different software. Thus, when merging SVs together, a recurrent approach considers the reciprocal overlap between the two coordinate ranges to assess whether these two SVs overlap sufficiently to be considered the same variant. How much overlap is sufficient is usually experimentally determined within each study: a classic approach consists in evaluating the trend of the number of merged SVs plotted against the overlap threshold, but, generally, 50 to 80% of reciprocal overlap is usually required to merge variants together. When confronted with the decision on how to combine our data, we had to solve the task twice: one first step involved the creation of a per-sample concordant set, hence merging the variants of different callers within each individual; after that, we could address the consensus set among all sample together to build up a cohort-level consensus dataset. To confound the decision on the direction to take to finally obtaining the set of variants to begin performing our analyses, lays the fact that the methods to define merging calls are not standardized among studies (Ho et al. 2019) (and references therein) and might differ depending on the information available to each project and their level of complexity. We used specifically designed software to merge our variants but encountered issues. Namely, the tested software do not account for intra-sample overlap, since they are designed to include files together, considering them two different individuals. This was not ideal because a first sample-level consensus list of CNVs had to be created in order to unite the strengths of our set of software. Merging the results of different callers together would have taken only the evaluation method used by the merging software, without taking into consideration the likelihood of contribution of each calling algorithm. Besides these issues, one tested software,

SURVIVOR (Jeffares et al. 2017) merges variants using relative distance between CNVs starts and ends, forcing the user to choose a fixed threshold distance to merge or not CNVs. We were not satisfied about the fixed distance threshold because, for a wide array of lengths and different configurations of pairs, it would have wrongfully excluded overlapping variants[21]. We also used other software, one in particular, FusorSV (Becker et al. 2018) sounded promising because of its implemented model comparing the query SVs set with known true SVs, with training phases refining the likelihood of merging true positive calls. Unfortunately, this has been one of the cases where difficulties in running the software and error messages could not be fully resolved even by contacting the authors, thus forcing us to choose another solution. As often reported in the literature, we also decided, despite our initial expectations, to develop a custom merging algorithm tailored to our dataset features and our need to resolve first in-sample consensus CNVs and then cohort-level merging.

For the first step, we wondered what to do when variants from two or more callers overlapped, as information about coordinates (starts and ends) and genotypes might not concur, in short: how do we choose? When a set of CNVs reciprocally overlapped 50% of their length, depending on the variant type, either deletion or duplication, we selected the coordinates and genotype of the most likely caller, mainly evaluated by an extensive research on SV software detection accuracy (Kosugi et al. 2019) thus combining the strengths of caller combinations. The following construction and implementation of the merging script had to face the difficulties of dealing with a complex data input, as every CNV had to be evaluated

[21] As an example, a threshold of 100bp, for two variants of ~200bp with 50bp distance between starts and ends would have been considered the same variant and merged; nevertheless, two ~10kbp variants having 150bp distance among starts and ends would have been excluded. The opposite scenario is also valid for larger thresholds.

against every other CNV from other callers. Restricting for CNV type and chromosome helped reduce this complexity which, nonetheless, had to be dealt with. Within the output results, for every merged CNV we retained the name of the callers providing the consensus coordinates and genotype respectively, to keep track of each caller contribution.

The subsequent cohort-level merging step proceeded in a similar way but without the need for the hierarchization of callers, as we already had, for each sample, one type of consensus information at each CNV. In this phase, the complexity multiplied manifold; indeed, thousands of CNVs had to be evaluated among all 138 samples. During the tests for the development of this part, we started suffering from excessive memory consumption (hundreds of Gb of RAM) as the creation of all these clusters of possible overlapping calls filled the available computer resources. To proceed, we had to make the code more efficient, refining the overlap evaluation approach, and strip down the data in input[22] allowing the code to work more effectively with this heavier load of data and perform as effortlessly as possible all the comparisons. Once we solved these issues, the merging step proceeded faster and with a significant decrease in memory consumption, making it feasible to work with the entire set of samples at once, this further refinement allowed us to obtain our long awaited dataset of consensus CNVs.

[22] We performed a sort of data compression, or transformation, where the input data was reduced and simplified to occupy as less memory as possible, while retaining the same unaltered level of information for subsequent evaluations.

4.2 Generating data and artificial complexity

One of the pillars of the scientific process is the necessity of replicating the results, namely being able to independently reproduce something that has already been done and presented. This reproducibility, however, does not bluntly pretend that a validation experiment will provide the same results as a previous iteration did, under whichever condition. Indeed, the verification process must be carried out in a specific controlled environment or with the same set of parameters under which the "parent" assay was performed with. Such condition is of paramount importance because maintaining the same conditions makes the two experiments comparable, otherwise it would not make sense even considering the second as a repetition of the former. Additionally, setting the same conditions helps attenuating the action of possible unknown variables which, at least in the case of a constant response to the known parameters, should affect the output in a proportional manner when manipulating the established parameters. Increasing the level of complexity of one experiment, with different variables involved in the production of a final result, might multiply exponentially the degree of interactions among them and among possible unexplored conditions.

Since its inception, whole genome sequencing has been standardized, mostly using two broad approaches in a few different platforms (Goodwin et al. 2016) and researchers developing this technology had to keep under control each variable to guarantee consistent results. Nonetheless, when using data from different sources, as in our project, the confounder action of batch effects may distort subsequent analyses. As discussed in the results, batch effect is the detection of differences among analysed groups that, overall, can be ascribed to technical properties rather than to actual intrinsic biological features. This phenomenon is known in WGS and, more broadly, is intrinsic to high-throughput experiments, where confounders may come

from processing groups, i.e., different protocols adopted to analyse the samples (Leek et al. 2010) which, for example, may result from laboratory-specific differences in microarray experiments (Irizarry et al. 2005). Additionally, also the date on which the experiments are carried out has been linked to variation among samples (Scherer 2009), nevertheless, these sources of differences are most likely only proxies masking the action of other players (namely reagent quality, ozone levels and temperature) and additional unknown factors may still have a role (Fare et al. 2003; Scherer 2009; Leek et al. 2010). Different aspects of whole genome sequencing can introduce batch effects, altering subsequent measurements; for example, the amount of dyes used to detect DNA bases may produce spurious G calls, different sequencing runs can introduce miscalibration of bases quality scores, varying read lengths can result in alignment errors (Lou and Therkildsen 2021). Ultimately, the repercussions of such complexity have also been identified in the 1KGP where batch effects have been observed, thus warning researchers of possible spurious results when preforming specific association studies and imputing variants (Anderson-Trocmé et al. 2020). As showed, our samples belong to three distinct projects that varied in sequencing location, read length, coverage and library preparation (PCR vs PCR-free) and when performing the first preliminary analyses, like PCA, an evident and strong separation of these three clusters highlighted the presence of batch effects in our dataset. To solve this issue, hindering the reliability of any further result we might get, we acted to thoroughly filter batch-affected variants. To do so, we started with a series of frequency-based filters, targeting singleton and rare variants, thought to be the most probable source of these batch effects afflicting our merged dataset. Subsequent filtering rounds and PCAs to assess the new situations proved us wrong, as the samples of the three Romani, SGDP and Mondal et al. samples kept clustering tightly together within the dataset but apart from each other among datasets. The effort to overcome this problem took a

considerable amount of time and research, contacting different researchers and revising literature of CNVs in population genetics. Concerning this last point, it was interesting to notice how the majority of the scrutinized studies often lack any reference to this batch effects problem and, frequently, this is because the analysed samples originate from a single source, one project generating the dataset to work with. Indeed, even if the studies actually took a population genetic overview of CNVs in different worldwide populations, they strictly maintained the analysis confined to either 1KGP, HGDP or others.

Eventually, the solution came from the combination of two actions: regenotyping and specific filtering the called CNVs. As discussed in the results, the filter step was carried out by following the pipeline of a dedicated software for regenotyping the called variants (GraphTyper2) while the second step included two phases: false positive and Hardy-Weinberg Equilibrium filtering respectively. As already anticipated at the beginning of the Discussion, due to the advancements in the research of SVs calling more recent studies tend to use dedicated software for genotyping cohort-level sets of merged SVs (Werling et al. 2018; Eggertsson et al. 2019; Larson et al. 2019; Almarri et al. 2020; Collins et al. 2020). After providing the coordinates of our calls and the reference genome for the GraphTyper2 pipeline and obtaining the software results, the output variants *per-se* still could not provide a solution to the batch effect issue. In fact, the output variants needed to be accurately filtered, according to specific parameters, in order to keep only high-quality regenotyped calls. We thus retained the correctly regenotyped variants and, with this new cleaned and more accurate dataset, proceeded with the subsequent two filters of the step two to finally obtain a more reliable consensus set of CNVs. Filtering steps are a routine practice in all population genetics studies: commonly used filters help retaining

informative and good quality SNPs and usually remove between 30-40% of the variants using arrays designed for population genetics studies, which increase to 64-80% when relying on general arrays or WGS approaches (Arauna et al. 2017; Font-Porterias et al. 2019; Bianco et al. 2020; Flores-Bello et al. 2021). In this project, the set of filters utilized to get to a refined high confidence callset (including the regenotyping step, as some called variants were reconsidered as homozygous for the reference, thus removing them from the original dataset) removed almost 70% of the initially called CNVs, thus greatly reducing the number of available data to begin the analyses with. Even considering the case where some filters might have been too conservative, thus removing actual *bona fide* variants, this reduction of CNVs suggests that many implementations for SV discovery in short-reads WGS still recover a great number of false positive calls, potentially coming also from the batch effects, and hampering subsequent analyses. This extensive regenotyping and filtering approach was indeed beneficial because it greatly attenuated the batch effects clustering previously seen using PCA. Concerning PCA, we could still see a mild pattern of general clusterization by dataset when comparing the first two principal components, which disappeared in PC3 and PC4, highlighting that some genotypes in the dataset might still be affected by batch effects and drive the differentiation among datasets. Nonetheless, other techniques of dimensionality reduction like UMAP, which is a non-linear method that acts better at representing the local structure of the different groups within a dataset, did show a more homogeneous clustering. PCA, on the other hand, using a linear data transformation, can be more prone to be biased by outliers. This points to the fact that the ways data is analysed and treated by different approaches may also emphasize specific aspects over others, influencing subsequent interpretations. Indeed, plotting the results of the PCA for PC3 and PC4, despite resulting in a general noisier sample distribution compared to the first two components, Romani appear to

cluster somehow in between Europeans and South Asians. UMAP on the other end, highlights more how Romani show tighter relationships with European samples; the former indeed cluster more closely to this group than to South Asian or Middle Eastern ones. Despite the difficulties caused by the influence of batch effects, the first analyses on our final dataset highlight the potential of CNVs, deletions in particular, to recapitulate the differentiation among populations, thus confirming previous indications from SNP-based research. Overall, these results, also reinforced by the subsequent admixture analysis showing a coherent apportionment of ancestry profiles among the analysed populations with knowledge accumulated so far, made us more confident about the sharp reduction of biased variants inherited from the joint analysis of different data sources.

4.3 CNVs and the Romani

For the first time in the scientific research covering the Romani population, we undertook an exploratory investigation of CNVs in this group providing an original and previously undescribed genome-wide perspective. The Romani attracted the attention of historians, anthropologists and geneticists but due to the historically overrepresentation of SNP studies (also considering the initial difficulties in studying SVs), thorough investigations utilizing CNVs have never been conducted in this minority ethnic group. As pinpointed through this work, CNVs constitute an important source of information, due to their implications in gene expression, diseases and evolutionary processes. Since Romani have not been addressed by whole genome studies until recently and CNVs gained more attention merely in the last decade, classically lagging behind in genetic research, we explored their variation finding intriguing results. We systematically identified more intronic deletions in genes showing intolerance to LoF in Romani than in other populations, which may be caused by a slight relaxation of purifying selection at these loci. We speculate that, while Romani had an intricate history of population size reduction, isolation and founder effects, admixture with encountered populations (as we saw, extensively with West Eurasian groups), representing a new source of variability, might have kept lower frequencies of more deleterious variants, while putatively slightly deleterious ones, as intronic deletions, might have been maintained either because of these dynamics in Romani history or because of the lower impact of these variants. Furthermore, using deletions, we manage to reconstruct the expected relationships among Romani, Eurasian, and South Asian samples, thus confirming the known predictive power of this type of copy number variation to inform over population ancestry and diversity.

Another result concerning CNVs is their taggability with SNPs, namely the patterns of LD we found for 78 deletions in linkage with 125 SNPs within

the GWAS catalogue. We discovered that the majority of tagged deletions mapped in regions with no genes, not residing within the gene of the SNP in linkage, and within the entire set of linked deletions are the closest to their relative SNP. This points either to the fact that, in a simplistic scenario, these intergenic deletions do not contribute to the disease trait associated with the SNP, or, conversely, being as close as ~70bp, they may alter the genomic environment around the SNP and thus have an influence over the trait. For Romani, a deletion in LD with a SNP associated with cholesterol is in line with previous reports of Romani individuals carrying higher frequencies of SNPs connected with high plasma cholesterol levels (Mendizabal et al. 2013). Another result concerning biomedical implications in Romani was an increased number of biological processes GO terms in LoF-intolerant genes containing intronic deletions, specifically for genes related to the functioning of the nervous system. Some of the different private disease-causing mutations of Romani also involve neurological disorders (Kalaydjieva et al. 2001; Morar et al. 2004; Kalaydjieva et al. 2005) and in general many genes of these sets of LoF-intolerant gene-harbouring deletions, overrepresented in different GO terms, are the same genes containing disease SNPs identified in Romani in a previous study (Mendizabal et al. 2013). The co-occurrence of SNPs and deletions in specific genes having related processes involved in biomedical traits, suggests that a level of interaction or relationship is not to be disregarded, particularly in light of the similar functions carried out by those genes, whose alteration generates similar pathologies.

In the present study, with the double intent to look for undetected information in Romani using CNVs and test this type of variant to infer their ability to recover evidence in an underrepresented population with a peculiar history, we gathered interesting insights on both sides. The higher load for putative slightly deleterious deletions in Romani, coherent with

previous studies attesting more frequent presence of deleterious SNPs (Mendizabal et al. 2013; Morar et al. 2013), adds up to the information concerning health-related markers collected so far using base-level point variations; corroborating while also integrating the insights about higher occurrences of those specific biomedical traits prevalent in Romani. The presence of deletions associated either with SNPs or within genes pertaining to conditions particularly present in Romani further highlights the importance of addressing CNVs in humans and specific populations (or minorities), where dissecting and characterizing the genetic contribution to diseases can increase the awareness of clinicians and favour more comprehensive and knowledgeable screening procedures.

On a more CNV-side, this study remarks the power of deletions as ancestry and diversity markers among populations, even when working with a limited dataset like the one we used post-filtering. Often disregarded, intronic variations and in particular SVs, might provide novel insights or deepen our understanding of more subtle levels of molecular dynamics acting on phenotype. The case of intronic deletions in genes particularly susceptible to mutation, more present in Romani, may signal the existence of additional refined layers of interactions between variants and clinically relevant traits and that such variants, usually removed in general populations, may still contribute to a number of conditions.

Throughout this thesis, we have stressed how most population genetic studies so far included mainly humans of European descent, perpetrating what is usually known as the Eurocentric bias. On the other hand, research trying to cover worldwide variation is usually affected by low numbers of samples in each category which, while acknowledged as an admirable undertaking, plays against a thorough resolution of underrepresented groups and keeps portraying an incomplete and biased picture of human genetic diversity. While initiatives to keep analysing more genomes around

the world exist, like projects addressing variation in Chinese people (Cao et al. 2020; Zhang et al. 2021), the DNA do Brasil project (Patrinos et al. 2020) or the H3Africa (Matovu et al. 2014), efforts will be necessary to constantly even out the unbalances in the amount of data gathered so far. Moreover, expanding the sample selection to small, isolated and underrepresented groups could reveal as a fruitful source of information. In addition, although the idea of deepening the knowledge on a growing number of human populations may sound intriguing for any researcher in the field, we must keep in mind the responsibilities we all have as first-line representatives of the scientific community. Notwithstanding the recognition of disparities in human population genetic research and the efforts to include additional information for a more diversified set of populations, researchers must question how this is achieved and to whom the benefits of such inclusion go. Shamefully, scientific misconduct in human genetics is not a novelty in the world of research, examples of lack of (or inadequate) consent forms to retrieve genetic data from indigenous communities in Africa and Australia or from Native Americans have been described (Claw et al. 2018a; Hudson et al. 2020) and while action is not taken to let the participants of indigenous communities to have a voice over the rights concerning their biological data, fostering their active participation and benefit from the studies, their involvement still quite resemble an exploitative action from whom conducts such projects (Claw et al. 2018a; Fox 2020; Hudson et al. 2020). Some glaring cases of misconduct, for example, concern the confirmed or alleged absence of proper consent forms for studies including ethnic minorities in China, using autosomal and Y-STR data, with the ability for the latter to match two DNA samples with an individual level precision, were retracted from the respective journals (Pan et al. 2020; Danyan et al. 2021; Normile 2021; Nothnagel et al. 2022). Other cases concern data collected decades ago, when attention towards correct sampling procedures and exhaustive project

explanations lacked even more than today. In this case, involving the Brazilian Karitiana indigenous people, their personal information have been used (exploited) repeatedly ever since (Munsterhjelm 2015). Extreme cases of racialized actions based on distorted biological justifications were also directed towards Romani communities in Slovakia, where Romani children represented up to ~90% of special school classes for children with "mild mental disabilities", whose existence and percentage composition were justified by the Slovakian government as being caused by the increased genetic disorders in Slovak Romani owing their higher inbreeding with respect to non-Romani Europeans (European Roma Rights Centre 2015; Open Society Foundations 2015; Amnesty International and European Roma Rights Centre 2017). Such reports badly stain the credibility and reputation of the scientific community to the public opinion and hinder future healthy collaborations between researchers and minority groups due to a legacy of known malpractice. In this context, the use of clear and rigorous consent forms, the involvement of representatives from the studied population and well-defined information about the privacy of the data are of primary importance if we want to keep advancing research based on solid grounds of trust and transparency (Hamel et al. 2016; Tiffin 2018; Tsosie et al. 2021). Another key point, which already has started to be addressed, is to put an end on the classic view of "us and them", further including communities in the studies (Claw et al. 2018b; Petraki 2020; Scelza et al. 2020) and actively provide tools for voluntary decisions for scientific research careers within members of indigenous communities, at risk of potentially being only seen more as the source of the data rather than an active component (https://singconsortium.org/). As a community of investigators, we should be able to amend past wrongdoings, also by actively question past, present and future actions and work for a truly open environment that can only benefit from the diverse spectrum of

backgrounds, ideas and heritage coming from different cultures joining to cooperate together.

4.4 Future perspectives

In the previous paragraph we briefly recapitulated the main results of this project and outlined how CNVs contribute to advancing the knowledge in this specific population. How could the dataset of called CNVs provide some additional information? What future directions it might point to and what insights may derive by analysing it from a different perspective?

An interesting question could be: since our investigation comprises different populations spanning a wide continental area and genetically diverse populations, including an isolate one, can we identify CNVs previously unreported in public databases? Addressing this question can highlight:

1. The benefits of studying copy number variation in underrepresented populations which, depending on their history and characteristics, have the potential to provide new insights for genomic features having either a population genetics or biomedical relevance, thus widening our understanding of the implications of CNVs and what dynamics co-occur in forming them.
2. How current CNV datasets, despite having a discreetly long history of data collection, still miss bits of information and hence are naturally incomplete, which in turn could foster further research targeting additional variation aiming to provide a more inclusive collection.
3. All those populations that still are less represented and not "CNV-covered", which as we saw should be, not surprisingly, mostly composed of populations of non-European descent.

To address these questions, a comprehensive dataset is needed as the foundation base the perform the analysis with. In this case, two major publicly available datasets: DGV and gnomAD (MacDonald et al. 2014;

Karczewski et al. 2020) could be a solid starting point to assess the overlap with deposited datasets, as these two repositories host together more than one million SVs.

To assess this possibility, we performed some initial test downloading the information of SV type and coordinates for the gnomAD dataset. A filtering step retained only those variants appearing in our dataset, namely deletions and duplications. To perform a convenient first assessment of the shared variants, we used the bed file provided by gnomAD and implemented the bedtools (Quinlan and Hall 2010) intersection methodology to obtain the common variants with our set. Bedtools implements an evaluation strategy where the user can request the software to only output those pairs of variants, in the same chromosome, reciprocally sharing a specific percentage of their length. To choose a reliable threshold of reciprocal overlap to use for the comparisons, we evaluated the number of the intersected variants (variants in common between the two sets) for every percentage point from 1 to 100. Plotting the number of intersected variants as a function of the threshold, one can evaluate specific threshold values where a marked change (reduction for example) of the analysed curve happens and thus use the corresponding threshold value as a reference for subsequent analyses. In our case we used an 80% threshold to perform the overlap. Doing so, we detected 516 deletions (16% of deletions in our dataset) and 189 duplications (39% of duplications in our dataset) that did not overlap with any of the available CNVs in gnomAD dataset. Romani and South Asian samples harbour the higher fraction of these new deletions, 17.8% and 24.2% respectively. Analysing the apportionment of these new variants among the populations in our dataset, most of new ones are in South Asian samples and early analyses within the populations represented in gnomAD supports a marked lower representation of this population in the dataset. This underrepresentation in gnomAD most probably accounts

for this skewed prevalence of unreported variants in our sample set from South Asia and thus points to the incomplete representation that such databases might provide to researchers. Additional validation is needed and the inclusion of the larger dataset form DGV; nonetheless, this first glance into the composition of current publicly available datasets for human SVs reflects accurately the research practices adopted so far.

6. REFERENCES

Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, et al. 2020. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* 583:83–89.

Abicht A, Stucka R, Karcagi V, Herczegfalvi A, Horváth R, Mortier W, Schara U, Ramaekers V, Jost W, Brunner J, et al. 1999. A common mutation (ε1267delG) in congenital myasthenic patients of Gypsy ethnic origin. *Neurology* [Internet] 53:1564–1564. Available from: https://n.neurology.org/content/53/7/1564

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21:974–984.

Achim V. 2004. The Roma in Romanian History. Central European University Press

Alharbi AF, Sheng N, Nicol K, Strömberg N, Hollox EJ. 2022. Balancing selection at the human salivary agglutinin gene (DMBT1) driven by host-microbe interactions. *iScience* 25:104189.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nature Reviews Genetics 2011 12:5* [Internet] 12:363–376. Available from: https://www.nature.com/articles/nrg2958

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics 2009 41:10* [Internet] 41:1061–1067. Available from: https://www.nature.com/articles/ng.437

Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, Chen Y, Hurles ME, Tyler-Smith C, Xue Y. 2020. Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell* 182:189-199.e15.

Alter G, Heckerman D, Schneidewind A, Fadda L, Kadie CM, Carlson JM, Oniangue-Ndza C, Martin M, Li B, Khakoo SI, et al. 2011. HIV-1 adaptation to NK-cell-mediated immune pressure. *Nature* 476:96–100.

Alter G, Martin MP, Teigen N, Carr WH, Suscovich TJ, Schneidewind A, Streeck H, Waring M, Meier A, Brander C, et al. 2007. Differential natural killer cell–mediated inhibition of HIV-1 replication based on distinct KIR/HLA subtypes. *Journal of Experimental Medicine* 204:3027–3036.

American Psychiatric Association. 2013. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association

Amnesty International, European Roma Rights Centre. 2017. A Lesson in Discrimination. Segregation of Romani Children in Primary Education in Slovakia.

Anderson ND, de Borja R, Young MD, Fuligni F, Rosic A, Roberts ND, Hajjar S, Layeghifard M, Novokmet A, Kowalski PE, et al. 2018. Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science (1979)* 361.

Anderson-Trocmé L, Farouni R, Bourgey M, Kamatani Y, Higasa K, Seo JS, Kim C, Matsuda F, Gravel S. 2020. Legacy Data Confound Genomics Studies. *Mol Biol Evol* [Internet] 37:2–10. Available from: https://academic.oup.com/mbe/article/37/1/2/5556817

Arauna LR, Mendoza-Revilla J, Mas-Sandoval A, Izaabel H, Bekada A, Benhamamouch S, Fadhlaoui-Zid K, Zalloua P, Hellenthal G, Comas D. 2017. Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa. *Mol Biol Evol* [Internet] 34:318–329. Available from: https://academic.oup.com/mbe/article/34/2/318/2680801

Arbogast T, Ouagazzal A-M, Chevalier C, Kopanitsa M, Afinowi N, Migliavacca E, Cowling BS, Birling M-C, Champy M-F, Reymond A, et al. 2016. Reciprocal Effects on Neurocognitive and Metabolic Phenotypes in Mouse Models of 16p11.2 Deletion and Duplication Syndromes. *PLoS Genet* 12:e1005709.

Arioka Y, Shishido E, Kushima I, Suzuki T, Saito R, Aiba A, Mori D, Ozaki N. 2021. Chromosome 22q11.2 deletion causes PERK-dependent vulnerability in dopaminergic neurons. *EBioMedicine* 63:103138.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AME, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the Major Structural Variant Alleles of the

Human Genome. *Cell* [Internet] 176:663-675.e19. Available from: http://dx.doi.org/10.1016/j.cell.2018.12.019

Azmanov DN, Dimitrova S, Florez L, Cherninkova S, Draganov D, Morar B, Saat R, Juan M, Arostegui JI, Ganguly S, et al. 2010. LTBP2 and CYP1B1 mutations and associated ocular phenotypes in the Roma/Gypsy founder population. *European Journal of Human Genetics 2011 19:3* [Internet] 19:326–333. Available from: https://www.nature.com/articles/ejhg2010181

Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al. 2013. Punctuated Evolution of Prostate Cancer Genomes. *Cell* 153:666–677.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* 7:552–564.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte R v., Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent Segmental Duplications in the Human Genome. *Science (1979)* 297:1003–1007.

Bakker P. 1995. Notes on the genesis of Caló and other Iberian Para-Romani varieties. *Romaní in contact with other languages*:125–150.

Bakker P. 2000. What is the Romani Language? University of Hertfordshire Press

Balachandran P, Beck CR. 2020. Structural variant identification and characterization. *Chromosome Research*.

Ballif BC, Sulpizio SG, Lloyd RM, Minier SL, Theisen A, Bejjani BA, Shaffer LG. 2007. The clinical utility of enhanced subtelomeric coverage in array CGH. *Am J Med Genet A* 143A:1850–1857.

Banci L, Camponeschi F, Ciofi-Baffoni S, Muzzioli R. 2015. Elucidating the Molecular Function of Human BOLA2 in GRX3-Dependent Anamorsin Maturation Pathway. *J Am Chem Soc* 137:16133–16143.

Bánfai Z, Ádám V, Pöstyéni E, Büki G, Czakó M, Miseta A, Melegh B. 2018. Revealing the impact of the Caucasus region on the genetic legacy of Romani people from genome-wide data. *PLoS One* [Internet] 13:e0202890. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0202890

Bank RuudA, Hettema EwaldH, Muijs MarianA, Pals G, Arwert F, Boomsma DorretI, Pronk JanC. 1992. Variation in gene copy number and polymorphism of the human salivary amylase isoenzyme system in Caucasians. *Hum Genet* 89.

Barany Z. 2001. The East European Gypsies in the imperial age. *Ethn Racial Stud* 24:50–63.

Barbour JD, Sriram U, Caillier SJ, Levy JA, Hecht FM, Oksenberg JR. 2007. Synergy or Independence? Deciphering the Interaction of HLA Class I and NK Cell KIR Alleles in Early HIV-1 Disease Progression. *PLoS Pathog* 3:e43.

Barth JMI, Berg PR, Jonsson PR, Bonanomi S, Corell H, Hemmer-Hansen J, Jakobsen KS, Johannesson K, Jorde PE, Knutsen H, et al. 2017. Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Mol Ecol* 26:4452–4466.

Bauman JGJ, Wiegant J, Borst P, van Duijn P. 1980. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Exp Cell Res* 128:485–490.

Becker T, Lee WP, Leone J, Zhu Q, Zhang C, Liu S, Sargent J, Shanker K, Mil-homens A, Cerveira E, et al. 2018. FusorSV: An algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol* 19:1–14.

Bernasovský I, Suchý J, Bernasovská K, Vargová T. 1976. Blood groups of Roms (Gypsies) in Czechoslovakia. *Am J Phys Anthropol* [Internet] 45:277–279. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ajpa.1330450213

Bi W, Park S-S, Shaw CJ, Withers MA, Patel PI, Lupski JR. 2003. Reciprocal Crossovers and a Positional Preference for Strand Exchange in Recombination Events Resulting in Deletion or Duplication of Chromosome 17p11.2. *The American Journal of Human Genetics* 73:1302–1315.

Bianco E, Laval G, Font-Porterias N, García-Fernández C, Dobon B, Sabido-Vera R, Sukarova Stefanovska E, Kučinskas V, Makukh H, Pamjav H, et al. 2020. Recent common origin, reduced population size, and marked admixture have shaped European roma genomes. *Mol Biol Evol* 37:3175–3187.

Bielski CM, Zehir A, Penson A v., Donoghue MTA, Chatila W, Armenia J, Chang MT, Schram AM, Jonsson P, Bandlamudi C, et al. 2018. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet* 50:1189–1195.

Blizinsky KD, Diaz-Castro B, Forrest MP, Schürmann B, Bach AP, Martin-de-Saavedra MD, Wang L, Csernansky JG, Duan J, Penzes P. 2016. Reversal of dendritic phenotypes in 16p11.2 microduplication mouse model neurons by pharmacological targeting of a network hub. *Proceedings of the National Academy of Sciences* 113:8520–8525.

Boehlke C, Zierau O, Hannig C. 2015. Salivary amylase – The enzyme of unspecialized euryphagous animals. *Arch Oral Biol* 60:1162–1176.

Boettger LM, Handsaker RE, Zody MC, Mccarroll SA. 2012. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nature Genetics 2012 44:8* [Internet] 44:881–885. Available from: https://www.nature.com/articles/ng.2334

Boyd-Bowman P. 1985. Índice geobiográfico de más de 56 mil pobladores de la América Hispánica. Instituto de Investigaciones Históricas, UNAM

Brás A, Rodrigues A, Rueff J. 2020. Copy number variations and constitutional chromothripsis (Review). *Biomed Rep*.

Brearley M. 2001. The persecution of Gypsies in Europe. *American Behavioral Scientist* 45:588–599.

Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T, Lalani SR, Graham B, Lee B, Shinawi M, et al. 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* 40:1466–1471.

Burssed B, Zamariolli M, Bellucco FT, Melaragno MI. 2022. Mechanisms of structural chromosomal rearrangement formation. *Mol Cytogenet* 15.

Bustamante CD, de La Vega FM, Burchard EG. 2011. Genomics for the world. *Nature* 475:163–165.

C Yuen RK, Merico D, Bookman M, L Howe J, Thiruvahindrapuram B, Patel R v, Whitney J, Deflaux N, Bingham J, Wang Z, et al. 2017.

Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci* 20:602–611.

Cabrera-Serrano M, Mavillard F, Biancalana V, Rivas E, Morar B, Hernández-Laín A, Olive M, Muelas N, Khan E, Carvajal A, et al. 2018. A Roma founder BIN1 mutation causes a novel phenotype of centronuclear myopathy with rigid spine. *Neurology* 91:e339–e348.

Calderon de Anda F, Rosario AL, Durak O, Tran T, Gräff J, Meletis K, Rei D, Soda T, Madabhushi R, Ginty DD, et al. 2012. Autism spectrum disorder susceptibility gene TAOK2 affects basal dendrite formation in the neocortex. *Nat Neurosci* 15:1022–1031.

Callen DF, Eyre H, Lane S, Shen Y, Hansmann I, Spinner N, Zackai E, McDonald-McGinn D, Schuffenhauer S, Wauters J. 1993. High resolution mapping of interstitial long arm deletions of chromosome 16: relationship to phenotype. *J Med Genet* 30:828–832.

Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, Xu Y, Du P, Wang T, Hu R, et al. 2020. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Research 2020 30:9* [Internet] 30:717–731. Available from: https://www.nature.com/articles/s41422-020-0322-9

Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 17:224–238.

Carver BS, Tran J, Gopalan A, Chen Z, Shaikh S, Carracedo A, Alimonti A, Nardella C, Varmeh S, Scardino PT, et al. 2009. Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. *Nat Genet* 41:619–624.

Caspersson T, Farber S, Foley GE, Kudynowski J, Modest EJ, Simonsson E, Wagh U, Zech L. 1968. Chemical differentiation along metaphase chromosomes. *Exp Cell Res* 49:219–222.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics* [Internet] 13:1–18. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-238

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al.

2014. Resolving the complexity of the human genome using single-molecule sequencing. *Nature 2014 517:7536* [Internet] 517:608–611. Available from: https://www.nature.com/articles/nature13907

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* 10:1–16.

Chaisson MJP, Wilson RK, Eichler EE. 2015. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* 16:627–640.

Chan S, Lam E, Saghbini M, Bocklandt S, Hastie A, Cao H, Holmlin E, Borodkin M. 2018. Structural Variation Detection and Analysis Using Bionano Optical Mapping. In: p. 193–203.

Charlson FJ, Ferrari AJ, Santomauro DF, Diminic S, Stockings E, Scott JG, McGrath JJ, Whiteford HA. 2018. Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophr Bull* 44:1195–1203.

Chen W, Hayward C, Wright AF, Hicks AA, Vitart V, Knott S, Wild SH, Pramstaller PP, Wilson JF, Rudan I, et al. 2011. Copy number variation across European populations. *PLoS One* 6.

Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E. 2003. Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. *Am J Hum Genet* [Internet] 73:285–300. Available from: http://www.cell.com/article/S0002929707619181/fulltext

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology 2009 4:4* [Internet] 4:265–270. Available from: https://www.nature.com/articles/nnano.2009.12

Claw KG, Anderson MZ, Begay RL, Tsosie KS, Fox K, Garrison NA, Bader ACC, Bardill J, Bolnick DAA, Brooks J, et al. 2018a. A framework for enhancing ethical genomic research with Indigenous communities. *Nature Communications 2018 9:1* [Internet] 9:1–7. Available from: https://www.nature.com/articles/s41467-018-05188-3

Claw KG, Anderson MZ, Begay RL, Tsosie KS, Fox K, Garrison NA, Bader ACC, Bardill J, Bolnick DAA, Brooks J, et al. 2018b. A framework for enhancing ethical genomic research with Indigenous communities. *Nature Communications 2018 9:1* [Internet] 9:1–7. Available from: https://www.nature.com/articles/s41467-018-05188-3

Collins FS, Brooks LD, Chakravarti A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* [Internet] 8:1229–1231. Available from: https://pubmed.ncbi.nlm.nih.gov/9872978/

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera A v., Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* 581:444–451.

Commissioner for Human Rights. 2012. Human rights of Roma and Travellers in Europe. Strasbourg

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. 2005. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics 2006 38:1* [Internet] 38:75–81. Available from: https://www.nature.com/articles/ng1697

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* [Internet] 464:704–712. Available from: http://dx.doi.org/10.1038/nature08516

Cook B, Wayne GF, Valentine A, Lessios A, Yeh E. 2013. Revisiting the evidence on health and health care disparities among the Roma: A systematic review 2003-2012. *Int J Public Health* [Internet] 58:885–911. Available from: https://link.springer.com/article/10.1007/s00038-013-0518-6

Cooper GM, Nickerson DA, Eichler EE. 2007. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* 39:S22–S29.

Cortés-Ciriano I, Lee JJ-K, Xi R, Jain D, Jung YL, Yang L, Gordenin D, Klimczak LJ, Zhang C-Z, Pellman DS, et al. 2020. Comprehensive

analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* 52:331–341.

Craig-Holmes AP, Shaw MW. 1971. Polymorphism of Human Constitutive Heterochromatin. *Science (1979)* 174:702–704.

Creanza N, Feldman MW. 2016. Worldwide genetic and cultural change in human evolution. *Curr Opin Genet Dev* 41:85–92.

Cui C, Shu W, Li P. 2016. Fluorescence In situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications. *Front Cell Dev Biol* 4.

Danyan Z, Cao G, Mingkun X, Xuejun C, Li X, Chenchen T, Ye Y. 2021. RETRACTED ARTICLE: Y Chromosomal STR haplotypes in Chinese Uyghur, Kazakh and Hui ethnic groups and genetic features of DYS448 null allele and DYS19 duplicated allele. *Int J Legal Med* 135:1119.

Davis LI, Blobel G. 1986. Identification and characterization of a nuclear pore complex protein. *Cell* 45:699–709.

Davy BE. 2003. Congenital hydrocephalus in hy3 mice is caused by a frameshift mutation in Hydin, a large novel gene. *Hum Mol Genet* 12:1163–1170.

Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, Yu K, Rubanova Y, Macintyre G, Demeulemeester J, et al. 2021. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* 184:2239-2254.e39.

Depienne C, Moreno-De-Luca D, Heron D, Bouteiller D, Gennetier A, Delorme R, Chaste P, Siffroi J-P, Chantot-Bastaraud S, Benyahia B, et al. 2009. Screening for Genomic Rearrangements and Methylation Abnormalities of the 15q11-q13 Region in Autism Spectrum Disorders. *Biol Psychiatry* 66:349–359.

Derenko M, Malyarchuk B, Bahmanimehr A, Denisova G, Perkova M, Farjadian S, Yepiskoposyan L. 2013. Complete Mitochondrial DNA Diversity in Iranians. *PLoS One* [Internet] 8:e80673. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0080673

De Rubeis S, Pasciuto E, Li KW, Fernández E, Di Marino D, Buzzi A, Ostroff LE, Klann E, Zwartkruis FJT, Komiyama NH, et al. 2013.

CYFIP1 Coordinates mRNA Translation and Cytoskeleton Remodeling to Ensure Proper Dendritic Spine Formation. *Neuron* 79:1169–1182.

Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators, Centers for Disease Control and Prevention (CDC). 2014. Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR Surveill Summ* 63:1–21.

Dhami P, Coffey AJ, Abbs S, Vermeesch JR, Dumanski JP, Woodward KJ, Andrews RM, Langford C, Vetrie D. 2005. Exon Array CGH: Detection of Copy-Number Changes at the Resolution of Individual Exons in the Human Genome. *The American Journal of Human Genetics* 76:750–762.

DiStefano C, Gulsrud A, Huberty S, Kasari C, Cook E, Reiter LT, Thibert R, Jeste SS. 2016. Identification of a distinct developmental and behavioral profile in children with Dup15q syndrome. *J Neurodev Disord* 8:19.

Dixson AF. 2012. Primate Sexuality. Oxford University Press

Dobon B, ter Horst R, Laayouni H, Mondal M, Bianco E, Comas D, Ioana M, Bosch E, Bertranpetit J, Netea MG. 2020. The shaping of immunological responses through natural selection after the Roma Diaspora. *Sci Rep* [Internet] 10:1–12. Available from: https://doi.org/10.1038/s41598-020-73182-1

Dobranici M, Buzea A, Popescu R. 2012. The cardiovascular risk factors of the Roma (Gypsies) people in Central-Eastern Europe: a review of the published literature. *J Med Life* [Internet] 5:382. Available from: /pmc/articles/PMC3565246/

Dobzhansky T. 1970. Genetics of the Evolutionary Process. Columbia University Press

Doggett NA, Xie G, Meincke LJ, Sutherland RD, Mundt MO, Berbari NS, Davy BE, Robinson ML, Rudd MK, Weber JL, et al. 2006. A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics* 88:762–771.

Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, Capdeville R, Talpaz M. 2001. Activity of a Specific Inhibitor of the

BCR-ABL Tyrosine Kinase in the Blast Crisis of Chronic Myeloid Leukemia and Acute Lymphoblastic Leukemia with the Philadelphia Chromosome. *New England Journal of Medicine* 344:1038–1042.

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* 17:1266–1277.

Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, Gudbjartsson DF, Stefansson K, Halldorsson B v., Melsted P. 2019. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun* [Internet] 10:1–8. Available from: http://dx.doi.org/10.1038/s41467-019-13341-9

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science (1979)* [Internet] 323:133–138. Available from: https://www.science.org/doi/10.1126/science.1162986

Elkalioubie A, Allorge D, Robriquet L, Wiart J-F, Garat A, Broly F, Fourrier F. 2011. Near-fatal tramadol cardiotoxicity in a CYP2D6 ultrarapid metabolizer. *Eur J Clin Pharmacol* 67:855–858.

Ena GF, Aizpurua-Iraola J, Font-Porterias N, Calafell F, Comas D. 2022. Population Genetics of the European Roma—A Review. *Genes (Basel)* 13:2068.

English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, Davis CF, Dahdouli M, Ma S, et al. 2015. Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics* [Internet] 16:1–15. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1479-3

English AC, Salerno WJ, Reid JG. 2014. PBHoney: Identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* [Internet] 15:1–7. Available from: https://link.springer.com/articles/10.1186/1471-2105-15-180

Escaramís G, Docampo E, Rabionet R. 2015. A decade of structural variants: Description, history and methods to detect structural variation. *Brief Funct Genomics* 14:305–314.

European Roma Rights Centre. 2015. Slovakia: Racist stereotyping should not determine education policy - International NGOs criticize Slovak Government. Available from: http://www.errc.org/press-releases/slovakia-racist-stereotyping-should-not-determine-education-policy--international-ngos-criticize-slovak-government

European Union Agency for Fundamental Rights. 2018. A persisting concern: Anti-Gypsyism as a barrier to Roma inclusion. Publications Office of the European Union Luxembourg

Ewing A, Meynert A, Churchman M, Grimes GR, Hollis RL, Herrington CS, Rye T, Bartos C, Croy I, Ferguson M, et al. 2021. Structural Variants at the BRCA1/2 Loci are a Common Source of Homologous Repair Deficiency in High-grade Serous Ovarian Carcinoma. *Clinical Cancer Research* 27:3201–3214.

Falchi M, El-Sayed Moustafa JS, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC, Sudmant PH, Dorajoo R, Al-Shafai MN, Bottolo L, et al. 2014. Low copy number of the salivary amylase gene predisposes to obesity. *Nat Genet* 46:492–497.

Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, Koch JE, LeProustt E, Marton MJ, Meyer MR, et al. 2003. Effects of atmospheric ozone on microarray data quality. *Anal Chem* 75:4672–4675.

Farré M, Micheletti D, Ruiz-Herrera A. 2013. Recombination Rates and Genomic Shuffling in Human and Chimpanzee—A New Twist in the Chromosomal Speciation Theory. *Mol Biol Evol* 30:853–864.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* 7:85–97.

Firth H v., Wright CF. 2011. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* 53:702–703.

Flint J, Hill AVS, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, Bana-Koiri J, Bhatia K, Alpers MP, Boyce AJ, et al. 1986. High frequencies of α-thalassaemia are the result of natural selection by malaria. *Nature* [Internet] 321:744–750. Available from: http://www.nature.com/articles/321744a0

Flores-Bello A, Bauduer F, Salaberria J, Oyharçabal B, Calafell F, Bertranpetit J, Quintana-Murci L, Comas D. 2021. Genetic origins, singularity, and heterogeneity of Basques. *Current Biology* 31:2167-2177.e4.

Földes ME, Covaci A. 2012. Research on Roma health and access to healthcare: State of the art and future challenges. *Int J Public Health* [Internet] 57:37–39. Available from: https://link.springer.com/article/10.1007/s00038-011-0312-2

Font-Porterias N, Arauna LR, Poveda A, Bianco E, Rebato E, Prata MJ, Calafell F, Comas D. 2019. European Roma groups show complex West Eurasian admixture footprints and a common South Asian genetic origin. *PLoS Genet* 15:e1008417.

Font-Porterias N, Caro-Consuegra R, Lucas-Sánchez M, Lopez M, Giménez A, Carballo-Mesa A, Bosch E, Calafell F, Quintana-Murci L, Comas D. 2021. The Counteracting Effects of Demography on Functional Genomic Variation: The Roma Paradigm. *Mol Biol Evol* 38:2804–2817.

Font-Porterias N, Giménez A, Carballo-Mesa A, Calafell F, Comas D. 2021. Admixture Has Shaped Romani Genetic Diversity in Clinically Relevant Variants. *Front Genet* 12:1–12.

Fox K. 2020. The Illusion of Inclusion — The "All of Us" Research Program and Indigenous Peoples' DNA. *New England Journal of Medicine* 383:411–413.

Francke U. 1994. Digitized and differentially shaded human chromosome ideograms for genomic applications. *Cytogenet Genome Res* [Internet] 65:206–219. Available from: https://www.karger.com/Article/FullText/133633

Fraser A. 1992. The gypsies. Oxford: Wiley-Blackwell

Fu R, Mokhtar SS, Phipps ME, Hoh BP, Xu S. 2018. A genome-wide characterization of copy number variations in native populations of Peninsular Malaysia. *European Journal of Human Genetics* [Internet] 26:886–897. Available from: http://dx.doi.org/10.1038/s41431-018-0120-8

Fu S, Wang A, Au KF. 2019. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol* [Internet] 20:1–17. Available from:

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1605-z

Fujita Y, Yamashita T. 2014. Axon growth inhibition by RhoA/ROCK in the central nervous system. *Front Neurosci* 8.

Fujiwara M, Yoshimoto T, Morita Y, Kamada M. 1992. Interstitial deletion of chromosome 16q: 16q22 is critical for 16q- syndrome. *Am J Med Genet* 43:561–564.

Gaudieri S, DeSantis D, McKinnon E, Moore C, Nolan D, Witt CS, Mallal SA, Christiansen FT. 2005. Killer immunoglobulin-like receptors and HLA act both independently and synergistically to modify HIV disease progression. *Genes Immun* 6:683–690.

Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, et al. 2011. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res* 21:1626–1639.

Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, Mitchell TJ, Rubanova Y, Anur P, Yu K, et al. 2020. The evolutionary history of 2,658 cancers. *Nature* 578:122–128.

Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH, Shafer N, Bernier R, Ferrero GB, Silengo M, et al. 2011. Relative Burden of Large CNVs on a Range of Neurodevelopmental Phenotypes. *PLoS Genet* 7:e1002334.

Gisselsson D, Pettersson L, Höglund M, Heidenblad M, Gorunova L, Wiegant J, Mertens F, Dal Cin P, Mitelman F, Mandahl N. 2000. Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proceedings of the National Academy of Sciences* 97:5357–5362.

Golzio C, Katsanis N. 2013. Genetic architecture of reciprocal CNVs. *Curr Opin Genet Dev* 23:240–248.

Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, Reymond A, Sun M, Sawa A, Gusella JF, et al. 2012. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* 485:363–367.

Gómez-Carballa A, Pardo-Seco J, Fachal L, Vega A, Cebey M, Martinón-Torres N, Martinón-Torres F, Salas A. 2013. Indian Signatures in the Westernmost Edge of the European Romani Diaspora: New Insight from Mitogenomes. *PLoS One* 8.

Gonzalez Castro LN, Wesseling P. 2021. The cIMPACT-NOW updates and their significance to current neuro-oncology practice. *Neurooncol Pract* 8:4–10.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, et al. 2005. The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. *Science (1979)* 307:1434–1440.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* [Internet] 17:333–351. Available from: http://dx.doi.org/10.1038/nrg.2016.49

Goossens M, Dozy AM, Embury SH, Zachariades Z, Hadjiminas MG, Stamatoyannopoulos G, Kan YW. 1980. Triplicated alpha-globin loci in humans. *Proceedings of the National Academy of Sciences* 77:518–521.

Gorthi A, Romero JC, Loranc E, Cao L, Lawrence LA, Goodale E, Iniguez AB, Bernard X, Masamsetti VP, Roston S, et al. 2018. EWS–FLI1 increases transcription to cause R-loops and block BRCA1 repair in Ewing sarcoma. *Nature* 555:387–391.

Greaves M. 2015. Evolutionary Determinants of Cancer. *Cancer Discov* 5:806–820.

Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, Angelicheva D, Calafell F, Oefner PJ, Shen P, et al. 2001. Origins and divergence of the Roma (gypsies). *Am J Hum Genet* [Internet] 69:1314–1331. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1235543&tool=pmcentrez&rendertype=abstract

Griswold AJ, Ma D, Cukier HN, Nations LD, Schmidt MA, Chung R-H, Jaworski JM, Salyakina D, Konidari I, Whitehead PL, et al. 2012. Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. *Hum Mol Genet* 21:3513–3523.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. *Pathogenetics* [Internet] 1:4. Available from: http://pathogeneticsjournal.biomedcentral.com/articles/10.1186/1755-8417-1-4

Gudbjartsson DF, Hicks A, Barnard J, Olafsdottir A, Gulcher JR, Frigge ML, Steinthorsdottir V, Ingason A, Stefansson H, Masson G, et al. 2005. A common inversion under selection in Europeans. *Nat Genet* 37:129–137.

Gusmão A, Gusmão L, Gomes V, Alves C, Calafell F, Amorim A, Prata MJ. 2008. A perspective on the history of the iberian gypsies provided by phylogeographic analysis of Y-chromosome lineages. *Ann Hum Genet* 72:215–227.

Hagenbüchle O, Bovey R, Young RA. 1980. Tissue-specific expression of mouse α-amylase genes: Nucleotide sequence of isoenzyme mRNAs from pancreas and salivary gland. *Cell* 21:179–187.

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC. 2010. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* [Internet] 26:1277–1283. Available from: https://academic.oup.com/bioinformatics/article/26/10/1277/194099

Hallmayer J. 2011. Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Arch Gen Psychiatry* 68:1095.

Hamdan A, Ewing A. 2022. Unravelling the tumour genome: The evolutionary and clinical impacts of structural variants in tumourigenesis. *Journal of Pathology* 257:479–493.

Hamel LM, Penner LA, Albrecht TL, Heath E, Gwede CK, Eggly S. 2016. Barriers to Clinical Trial Enrollment in Racial and Ethnic Minority Patients with Cancer. *http://dx.doi.org/10.1177/107327481602300404* [Internet] 23:327–337. Available from: https://journals.sagepub.com/doi/abs/10.1177/107327481602300404

Han M v., Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. *Genome Res* 19:859–867.

Hancock I. 1984. Romani and Angloromani. *Language in the British Isles, ed. P. Trudgill*:367–383.

Hancock I. 2007. On the Interpretation of a Word:'porrajmos' as holocaust. In: Travellers, gypsies, Roma: The demonisation of difference. Vol. 53.

Hancock IF. 2002. We are the Romani people. Univ of Hertfordshire Press

Harel T, Lupski JR. 2018. Genomic disorders 20 years on-mechanisms for clinical manifestations. *Clin Genet* 93:439–449.

Harvard C, Malenfant P, Koochek M, Creighton S, Mickelson E, Holden J, Lewis M, Rajcan-Separovic E. 2005. A variant Cri du Chat phenotype and autism spectrum disorder in a subject with de novo cryptic microdeletions involving 5p15.2 and 3p24.3-25 detected using whole genomic array CGH. *Clin Genet* 67:341–351.

Hastings PJ, Ira G, Lupski JR. 2009. A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. *PLoS Genet* 5:e1000327.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* 10:551–564.

Hattori A, Fukami M. 2020. Established and Novel Mechanisms Leading to de novo Genomic Rearrangements in the Human Germline. *Cytogenet Genome Res* 160:167–176.

Haunhorst P, Hanschmann E-M, Bräutigam L, Stehling O, Hoffmann B, Mühlenhoff U, Lill R, Berndt C, Lillig CH. 2013. Crucial function of vertebrate glutaredoxin 3 (PICOT) in iron homeostasis and hemoglobin maturation. *Mol Biol Cell* 24:1895–1903.

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016a. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* 7:1–10.

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016b. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature*

Communications 2016 7:1 [Internet] 7:1–10. Available from: https://www.nature.com/articles/ncomms12989

Heim MH, Meyer UA. 1992. Evolution of a highly polymorphic human cytochrome P450 gene cluster: CYP2D6. *Genomics* 14:49–58.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. *Science* [Internet] 343:747. Available from: /pmc/articles/PMC4209567/

Higuchi R, Iwane T, Iida A, Nakajima K. 2020. Copy Number Variation of the Salivary Amylase Gene and Glucose Metabolism in Healthy Young Japanese Women. *J Clin Med Res* 12:184–189.

Hilker R, Helenius D, Fagerlund B, Skytthe A, Christensen K, Werge TM, Nordentoft M, Glenthøj B. 2018. Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. *Biol Psychiatry* 83:492–498.

Ho SS, Urban AE, Mills RE. 2019. Structural variation in the sequencing era. *Nat Rev Genet* [Internet]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31729472

Hochhaus A, Baccarani M, Silver RT, Schiffer C, Apperley JF, Cervantes F, Clark RE, Cortes JE, Deininger MW, Guilhot F, et al. 2020a. European LeukemiaNet 2020 recommendations for treating chronic myeloid leukemia. *Leukemia* 34:966–984.

Hochhaus A, Baccarani M, Silver RT, Schiffer C, Apperley JF, Cervantes F, Clark RE, Cortes JE, Deininger MW, Guilhot F, et al. 2020b. European LeukemiaNet 2020 recommendations for treating chronic myeloid leukemia. *Leukemia* 34:966–984.

Hochhaus A, Larson RA, Guilhot F, Radich JP, Branford S, Hughes TP, Baccarani M, Deininger MW, Cervantes F, Fujihara S, et al. 2017. Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid Leukemia. *New England Journal of Medicine* 376:917–927.

Holland AJ, Cleveland DW. 2009. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat Rev Mol Cell Biol* 10:478–487.

Holland AJ, Cleveland DW. 2012. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal

rearrangements. *Nature Medicine 2012 18:11* [Internet] 18:1630–1638. Available from: https://www.nature.com/articles/nm.2988

Horev G, Ellegood J, Lerch JP, Son Y-EE, Muthuswamy L, Vogel H, Krieger AM, Buja A, Henkelman RM, Wigler M, et al. 2011. Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proceedings of the National Academy of Sciences* 108:17076–17081.

Hsieh PH, Dang V, Vollger MR, Mao Y, Huang TH, Dishuck PC, Baker C, Cantsilieris S, Lewis AP, Munson KM, et al. 2021. Evidence for opposing selective forces operating on human-specific duplicated TCAF genes in Neanderthals and humans. *Nat Commun* [Internet] 12:1–14. Available from: http://dx.doi.org/10.1038/s41467-021-25435-4

Hsieh PH, Vollger MR, Dang V, Porubsky D, Baker C, Cantsilieris S, Hoekzema K, Lewis AP, Munson KM, Sorensen M, et al. 2019. Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science (1979)* [Internet] 366. Available from: https://www.science.org/doi/10.1126/science.aax2083

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2017a. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* [Internet] 27:677–685. Available from: https://genome.cshlp.org/content/27/5/677.full

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2017b. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* [Internet] 27:677–685. Available from: https://genome.cshlp.org/content/27/5/677.full

Hudson M, Garrison NA, Sterling R, Caron NR, Fox K, Yracheta J, Anderson J, Wilcox P, Arbour L, Brown A, et al. 2020. Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nature Reviews Genetics 2020 21:6* [Internet] 21:377–384. Available from: https://www.nature.com/articles/s41576-020-0228-x

Hurles ME, Dermitzakis ET, Tyler-Smith C. 2008. The functional impact of structural variation in humans. *Trends in Genetics* 24:238–245.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004a. Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004b. Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951.

Ionita-Laza I, Xu B, Makarov V, Buxbaum JD, Roos JL, Gogos JA, Karayiorgou M. 2014. Scan statistic-based analysis of exome sequencing data identifies *FAN1* at 15q13.3 as a susceptibility gene for schizophrenia and autism. *Proceedings of the National Academy of Sciences* 111:343–348.

Iovita Radu P., Schurr TG. 2004. Reconstructing the Origins and Migrations of Diasporic Populations: The Case of the European Gypsies. *Am Anthropol* [Internet] 106:267–281. Available from: https://onlinelibrary.wiley.com/doi/full/10.1525/aa.2004.106.2.267

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JGN, Geoghegan J, Germino G, et al. 2005. Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2:345–349.

Iskow RC, Gokcumen O, Lee C. 2012. Exploring the role of copy number variants in human adaptation. *Trends in Genetics* [Internet] 28:245–257. Available from: http://dx.doi.org/10.1016/j.tig.2012.03.002

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, et al. 2008. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84:148–161.

Jacquemont M-L, Sanlaville D, Redon R, Raoul O, Cormier-Daire V, Lyonnet S, Amiel J, le Merrer M, Heron D, de Blois M-C, et al. 2006. Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *J Med Genet* 43:843–849.

Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, Martinet D, Shen Y, Valsesia A, Beckmann ND, et al.

2011. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 478:97–102.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel K v, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* 36:321–323.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications 2017 8:1* [Internet] 8:1–11. Available from: https://www.nature.com/articles/ncomms14061

Johansson I, Lundqvist E, Bertilsson L, Dahl ML, Sjöqvist F, Ingelman-Sundberg M. 1993. Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proceedings of the National Academy of Sciences* 90:11825–11829.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001a. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413:514–519.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001b. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413:514–519.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484:55–61.

Kalaydjieva L, Gresham D, Calafell F. 2001. Genetic studies of the Roma (Gypsies): A review. *BMC Med Genet* 2.

Kalaydjieva L, Hallmayer J, Chandler D, Savov A, Nikolova A, Angelicheva D, King RHH, Ishpekova B, Honeyman K, Calafell F, et al. 1996. Gene mapping in Gypsies identifies a novel demyelinating neuropathy on chromosome 8q24. *Nat Genet* [Internet] 14:214–217. Available from: https://pubmed.ncbi.nlm.nih.gov/8841199/

Kalaydjieva L, Morar B, Chaix R, Tang H. 2005. A newly discovered founder population: The Roma/Gypsies. *BioEssays* 27:1084–1094.

Kallioniemi A, Kallioniemi O-P, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. 1992. Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. *Science (1979)* 258:818–821.

Kan YW, Dozy AM, Varmus HE, Taylor JM, Holland JP, Lie-Injo LE, Ganesan J, Todd D. 1975. Deletion of α-globin genes in haemoglobin-H disease demonstrates multiple α-globin structural loci. *Nature 1975 255:5505* [Internet] 255:255–256. Available from: https://www.nature.com/articles/255255a0

Kanduri C, Ukkola-Vuoti L, Oikkonen J, Buck G, Blancher C, Raijas P, Karma K, Lähdesmäki H, Järvelä I. 2013. The genome-wide landscape of copy number variations in the MUSGEN study provides evidence for a founder effect in the isolated Finnish population. *European Journal of Human Genetics 2013 21:12* [Internet] 21:1411–1416. Available from: https://www.nature.com/articles/ejhg201360

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581:434–443.

Kato H, Kimura H, Kushima I, Takahashi N, Aleksic B, Ozaki N. 2022. The genetic architecture of schizophrenia: review of large-scale genetic studies. *J Hum Genet*.

Kaushik Tiwari M, Colon-Rios DA, Tumu HCR, Liu Y, Quijano E, Krysztofiak A, Chan C, Song E, Braddock DT, Suh H-W, et al. 2022. Direct targeting of amplified gene loci for proapoptotic anticancer therapy. *Nat Biotechnol* 40:325–334.

Kawashima A, Satta Y. 2014. Substrate-Dependent Evolution of Cytochrome P450: Rapid Turnover of the Detoxification-Type and Conservation of the Biosynthesis-Type. *PLoS One* 9:e100059.

Kenrick D. 2004. Gypsies: from the Ganges to the Thames. Univ of Hertfordshire Press

Kenrick D. 2007. Historical dictionary of the Gypsies (Romanies). Scarecrow Press

Kimura M, others. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.

Kimura S, Umeno M, Skoda RC, Meyer UA, Gonzalez FJ. 1989. The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic CYP2D6 gene, a related gene, and a pseudogene. *Am J Hum Genet* 45:889–904.

Kircher M, Kelso J. 2010. High-throughput DNA sequencing – concepts and limitations. *BioEssays* [Internet] 32:524–536. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/bies.200900181

Kirkpatrick M, Barton N. 2006. Chromosome Inversions, Local Adaptation and Speciation. *Genetics* 173:419–434.

Kirov G, Rees E, Walters JTR, Escott-Price V, Georgieva L, Richards AL, Chambert KD, Davies G, Legge SE, Moran JL, et al. 2014. The Penetrance of Copy Number Variations for Schizophrenia and Developmental Delay. *Biol Psychiatry* 75:378–385.

Klarić IM, Salihović MP, Lauc LB, Zhivotovsky LA, Rootsi S, Janićijević B. 2009. Dissecting the molecular architecture and origin of Bayash Romani patrilineages: Genetic influences from South-Asia and the Balkans. *Am J Phys Anthropol* 138:333–342.

Kloosterman WP, Francioli LC, Hormozdiari F, Marschall T, Hehir-Kwa JY, Abdellaoui A, Lameijer EW, Moed MH, Koval V, Renkens I, et al. 2015. Characteristics of de novo structural changes in the human genome. *Genome Res* [Internet] 25:792–801. Available from: https://genome.cshlp.org/content/25/6/792.full

Kochanowski J. 1968. Black Gypsies, White Gypsies: The Gypsies within the perspective of Indo-European Migrations. *Diogenes* 16:27–47.

Kohmoto T, Naruto T, Watanabe M, Fujita Y, Ujiro S, Okamoto N, Horikawa H, Masuda K, Imoto I. 2017. A 590 kb deletion caused by non-allelic homologous recombination between two LINE-1 elements in a patient with mesomelia-synostosis syndrome. *Am J Med Genet A* 173:1082–1086.

Korbel Jan O., Campbell PJ. 2013. Criteria for Inference of Chromothripsis in Cancer Genomes. *Cell* 152:1226–1236.

Korbel Jan O, Campbell PJ. 2013. Criteria for inference of chromothripsis in cancer genomes. *Cell* 152:1226–1236.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science (1979)* [Internet] 318:420–426. Available from: https://www.science.org/doi/10.1126/science.1149504

Korbel JO, Urban AE, Grubert F, Du J, Royce TE, Starr P, Zhong G, Emanuel BS, Weissman SM, Snyder M, et al. 2007. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci U S A* [Internet] 104:10110–10115. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0703834104

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* 20:8–11.

Krimbas CB, Powell JR. 1992. Drosophila inversion polymorphism. CRC Press

Kruglyak L, Nickerson DA. 2001. Variation is the spice of life. *Nat Genet* 27:234–236.

Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, Gilliam TC, Nowak NJ, Cook EH, Dobyns WB, et al. 2007. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* 17:628–638.

Küpper C, Stocks M, Risse JE, dos Remedios N, Farrell LL, McRae SB, Morgan TC, Karlionova N, Pinchuk P, Verkuil YI, et al. 2016. A supergene determines highly divergent male reproductive morphs in the ruff. *Nat Genet* 48:79–83.

Kushima I, Aleksic B, Nakatochi M, Shimamura T, Shiino T, Yoshimi A, Kimura H, Takasaki Y, Wang C, Xing J, et al. 2017. High-resolution copy number variation analysis of schizophrenia in Japan. *Mol Psychiatry* 22:430–440.

LaFramboise T. 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* [Internet] 37:4181. Available from: /pmc/articles/PMC2715261/

Lai WR, Johnson MD, Kucherlapati R, Park PJ. 2005. Comparative analysis of algorithms for identifying amplifications and deletions in

array CGH data. *Bioinformatics* [Internet] 21:3763–3770. Available from: https://pubmed.ncbi.nlm.nih.gov/16081473/

Lam HYK, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, et al. 2012. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nature Biotechnology 2012 30:3* [Internet] 30:226–229. Available from: https://www.nature.com/articles/nbt.2134

Lam KWG, Jeffreys AJ. 2007. Processes of de novo duplication of human α-globin genes. *Proceedings of the National Academy of Sciences* [Internet] 104:10950–10955. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0703856104

Lappalainen T, Scott AJ, Brandt M, Hall IM. 2019. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* 177:70–84.

Larson DE, Abel HJ, Chiang C, Badve A, Das I, Eldred JM, Layer RM, Hall IM. 2019. svtools: population-scale analysis of structural variation. *Bioinformatics* [Internet] 35:4782–4787. Available from: https://academic.oup.com/bioinformatics/article/35/22/4782/5520944

Lasa A, Piccolo F, de Diego C, Jeanpierre M, Colomer J, Rodríguez MJ, Urtizberea JA, Baiget M, Kaplan JC, Gallano P. 1998. Severe limb girdle muscular dystrophy in Spanish gypsies: further evidence for a founder mutation in the γ-sarcoglycan gene. *European Journal of Human Genetics 1998 6:4* [Internet] 6:396–399. Available from: https://www.nature.com/articles/5200197

Lau Y-L, Chan L-C, Chan Y-YA, Ha S-Y, Yeung C-Y, Waye JS, Chui DHK. 1997. Prevalence and Genotypes of α- and β-Thalassemia Carriers in Hong Kong — Implications for Population Screening. *New England Journal of Medicine* 336:1298–1301.

Lee H, Schatz MC. 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* [Internet] 28:2097–2105. Available from: https://academic.oup.com/bioinformatics/article/28/16/2097/323484

Lee JA, Carvalho CMB, Lupski JR. 2007a. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell* 131:1235–1247.

Lee JA, Carvalho CMB, Lupski JR. 2007b. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell* 131:1235–1247.

Lee JJ-K, Park S, Park H, Kim S, Lee Jongkeun, Lee Junehawk, Youk J, Yi K, An Y, Park IK, et al. 2019. Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma. *Cell* 177:1842-1857.e21.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11:733–739.

Leffler EM, Band G, Busby GBJ, Kivinen K, Le QS, Clarke GM, Bojang KA, Conway DJ, Jallow M, Sisay-Joof F, et al. 2017. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science (1979)* [Internet] 356:1140–1152. Available from: https://www.science.org/doi/10.1126/science.aam6393

Lehrman MA, Schneider WJ, Südhof TC, Brown MS, Goldstein JL, Russell DW. 1985. Mutation in LDL Receptor: Alu-Alu Recombination Deletes Exons Encoding Transmembrane and Cytoplasmic Domains. *Science (1979)* 227:140–146.

Leija-Salazar M, Sedlazeck FJ, Toffoli M, Mullin S, Mokretar K, Athanasopoulou M, Donald A, Sharma R, Hughes D, Schapira AHV, et al. 2019. Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. *Mol Genet Genomic Med* 7:e564.

Lek M, Karczewski KJ, Minikel E v., Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature 2016 536:7616* [Internet] 536:285–291. Available from: https://www.nature.com/articles/nature19057

Leppa VM, Kravitz SN, Martin CL, Andrieux J, Le Caignec C, Martin-Coignard D, DyBuncio C, Sanders SJ, Lowe JK, Cantor RM, et al. 2016. Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families. *The American Journal of Human Genetics* 99:540–554.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The Diploid Genome Sequence of an Individual Human. *PLoS Biol* 5:e254.

Lewy G. 2000. The Nazi persecution of the Gypsies. Oxford University Press

Li Q-H, Wang Y-Z, Tu J, Liu C-W, Yuan Y-J, Lin R, He W-L, Cai S-R, He Y-L, Ye J-N. 2020. Anti-EGFR therapy in metastatic colorectal cancer: mechanisms and potential regimens of drug resistance. *Gastroenterol Rep (Oxf)* 8:179–191.

Li X, Tan L, Liu X, Lei S, Yang T, Chen X, Zhang F, Fang Y, Guo Y, Zhang Liang, et al. 2010. A genome wide association study between copy number variation (CNV) and human height in Chinese population. *Journal of Genetics and Genomics* 37:779–785.

Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE, et al. 2020. Patterns of somatic structural variation in human cancer genomes. *Nature* 578:112–121.

Lieber MR. 2010. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. *Annu Rev Biochem* 79:181–211.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (1979)* [Internet] 326:289–293. Available from: https://www.science.org/doi/10.1126/science.1181369

Lima L, Marchet C, Caboche S, da Silva C, Istace B, Aury J-M, Touzet H, Chikhi R. 2020. Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data. *Brief Bioinform* 21:1164–1181.

Linardopoulou E v., Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437:94–100.

Lindsay SJ, Khajavi M, Lupski JR, Hurles ME. 2006. A Chromosomal Rearrangement Hotspot Can Be Identified from Population Genetic Variation and Is Coincident with a Hotspot for Allelic

Recombination. *The American Journal of Human Genetics* 79:890–902.

Liu M, Rehman S, Tang X, Gu K, Fan Q, Chen D, Ma W. 2019. Methodologies for Improving HDR Efficiency. *Front Genet* 9.

Liu P, Erez A, Nagamani SCS, Dhar SU, Kołodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, et al. 2011. Chromosome Catastrophes Involve Replication Mechanisms Generating Complex Genomic Rearrangements. *Cell* 146:889–903.

Liu S, Yao L, Ding D, Zhu H. 2010. CCL3L1 Copy Number Variation and Susceptibility to HIV-1 Infection: A Meta-Analysis. *PLoS One* 5:e15778.

Loftus BJ, Kim U-J, Sneddon VP, Kalush F, Brandon R, Fuhrmann J, Mason T, Crosby ML, Barnstead M, Cronin L, et al. 1999. Genome Duplications and Other Features in 12 Mb of DNA Sequence from Human Chromosome 16p and 16q. *Genomics* 60:295–308.

Loimaranta V, Jakubovics NS, Hytönen J, Finne J, Jenkinson HF, Strömberg N. 2005. Fluid- or Surface-Phase Human Salivary Scavenger Protein gp340 Exposes Different Bacterial Recognition Properties. *Infect Immun* 73:2245–2252.

Long BR, Ndhlovu LC, Oksenberg JR, Lanier LL, Hecht FM, Nixon DF, Barbour JD. 2008. Conferral of Enhanced Natural Killer Cell Function by KIR3DS1 in Early Human Immunodeficiency Virus Type 1 Infection. *J Virol* 82:4785–4792.

Lord C, Brugha TS, Charman T, Cusack J, Dumas G, Frazier T, Jones EJH, Jones RM, Pickles A, State MW, et al. 2020. Autism spectrum disorder. *Nat Rev Dis Primers* 6:5.

Lou H, Li S, Jin W, Fu R, Lu D, Pan X, Zhou H, Ping Y, Jin L, Xu S. 2015. Copy number variations and genetic admixtures in three Xinjiang ethnic minority groups. *European Journal of Human Genetics* 23:536–542.

Lou RN, Therkildsen NO. 2021. Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation. *Mol Ecol Resour*.

Louzada S, Algady W, Weyell E, Zuccherato LW, Brajer P, Almalki F, Scliar MO, Naslavsky MS, Yamamoto GL, Duarte YAO, et al. 2020. Structural variation of the malaria-associated human glycophorin A-B-E region. *BMC Genomics* [Internet] 21:1–16. Available from: https://link.springer.com/articles/10.1186/s12864-020-06849-8

Lowry DB, Willis JH. 2010. A Widespread Chromosomal Inversion Polymorphism Contributes to a Major Life-History Transition, Local Adaptation, and Reproductive Isolation. *PLoS Biol* 8:e1000500.

Lutz BD. 1995. Gypsies as victims of the holocaust. *Holocaust Genocide Stud* 9:346–359.

Ly P, Cleveland DW. 2017. Rebuilding Chromosomes After Catastrophe: Emerging Mechanisms of Chromothripsis. *Trends Cell Biol* 27:917–930.

MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res* 42:986–992.

Mack H, Swisshelm K. 2013. Robertsonian Translocations. In: Brenner's Encyclopedia of Genetics: Second Edition. Elsevier Inc. p. 301–305.

MacKinnon RN, Campbell LJ. 2013. Chromothripsis under the microscope: a cytogenetic perspective of two cases of AML with catastrophic chromosome rearrangement. *Cancer Genet* 206:238–251.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: The long and the short of it. *Genome Biol* 20:1–14.

Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, Antaki D, Shetty A, Holmans PA, Pinto D, et al. 2017. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* 49:27–35.

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, et al. 2008. Structural Variation of Chromosomes in Autism Spectrum Disorder. *The American Journal of Human Genetics* 82:477–488.

Martin MP, Bashirova A, Traherne J, Trowsdale J, Carrington M. 2003. Cutting Edge: Expansion of the KIR Locus by Unequal Crossing Over. *The Journal of Immunology* 171:2192–2195.

Martínez-Cruz B, Mendizabal I, Harmant C, de Pablo R, Ioana M, Angelicheva D, Kouvatsi A, Makukh H, Netea MG, Pamjav H, et al. 2016. Origins, admixture and founder lineages in European Roma. *European Journal of Human Genetics* 24:937–943.

Marushiakova E, Popov VZ, Popov V. 2001. Gypsies in the Ottoman Empire: A Contribution to the History of the Balkans. Univ of Hertfordshire Press

Matovu E, Bucheton B, Chisi J, Enyaru J, Hertz-Fowler C, Koffi M, Macleod A, Mumba D, Sidibe I, Simo G, et al. 2014. Enabling the genomic revolution in Africa. *Science (1979)* 344:1346–1348.

Matras Y. 2002. Romani: A linguistic introduction. Cambridge University Press

Matras Y, Bakker P, K͡iuchukov Khristo. 1997. The Typology and Dialectology of Romani. *The Typology and Dialectology of Romani*:1–255.

Matras Y, Gardner H, Jones C, Schulman V. 2007. Angloromani: A Different Kind of Language? *Anthropological Linguistics* [Internet] 49:142–184. Available from: http://www.jstor.org/stable/27641824

May J, Evans JA, Timmann C, Ehmen C, Busch W, Thye T, Agbenyega T, Horstmann RD. 2007. Hemoglobin Variants and Disease Manifestations in Severe Falciparum Malaria. *JAMA* 297:2220.

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. 2006. Common deletion polymorphisms in the human genome. *Nat Genet* 38:86–92.

McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, et al. 2009. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 41:1223–1227.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471:216–219.

Melegh BI, Banfai Z, Hadzsiev K, Miseta A, Melegh B. 2017. Refining the South Asian Origin of the Romani people. *BMC Genet* 18:1–13.

Mendizabal I, Lao O, Marigorta UM, Kayser M, Comas D. 2013. Implications of Population History of European Romani on Genetic Susceptibility to Disease. *Hum Hered* [Internet] 76:194–200. Available from: https://www.karger.com/Article/FullText/360762

Mendizabal I, Lao O, Marigorta UM, Wollstein A, Gusmão L, Ferak V, Ioana M, Jordanova A, Kaneva R, Kouvatsi A, et al. 2012. Reconstructing the population history of European Romani from genome-wide data. *Current Biology* 22:2342–2349.

Mendizabal I, Valente C, Gusmão A, Alves C, Gomes V, Goios A, Parson W, Calafell F, Alvarez L, Amorim A, et al. 2011. Reconstructing the Indian origin and dispersal of the european Roma: A maternal genetic perspective. *PLoS One* 6:1–10.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65.

Minárik G, Ferák V, Feráková E, Ficek A, Poláková H, Kádasi Ľ. 2003. High Frequency of GJB2 Mutation W24X among Slovak Romany (Gypsy) Patients with Non-Syndromic Hearing Loss (NSHL). *Gen. Physiol. Biophys* [Internet] 22:549–556. Available from: http://www.crg.es/deafness/

Mirkin E v., Mirkin SM. 2007. Replication Fork Stalling at Natural Impediments. *Microbiology and Molecular Biology Reviews* 71:13–35.

Mitchell TJ, Turajlic S, Rowan A, Nicol D, Farmery JHR, O'Brien T, Martincorena I, Tarpey P, Angelopoulos N, Yates LR, et al. 2018. Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell* 173:611-623.e17.

Mohamad Isa II, Jamaluddin J, Achim NH, Abubakar S. 2020. Population-specific profiling of CCL3L1 copy number of the three major ethnic groups in Malaysia and the implication on HIV susceptibility. *Gene* 754:144821.

Mollenhauer J, Wiemann S, Scheurlen W, Korn B, Hayashi Y, Wilgenbus KK, von Deimling A, Poustka A. 1997. DMBT1, a new member of

the SRCR superfamily, on chromosome 10q25.3–26.1 is deleted in malignant brain tumours. *Nat Genet* 17:32–39.

Moorjani P, Patterson N, Loh PR, Lipson M, Kisfali P, Melegh BI, Bonin M, Kádaši Ľ, Rieß O, Berger B, et al. 2013. Reconstructing Roma History from Genome-Wide Data. *PLoS One* 8.

Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D, Singh L. 2013. Genetic Evidence for Recent Population Mixture in India. *The American Journal of Human Genetics* 93:422–438.

Morar B, Azmanov DN, Kalaydjieva L. 2013. Roma (Gypsies): Genetic Studies. *eLS* [Internet]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/9780470015902.a0006239.pub3

Morar B, Gresham D, Angelicheva D, Tournev I, Gooding R, Guergueltcheva V, Schmidt C, Abicht A, Lochmuller H, Tordai A, et al. 2004. Mutation history of the roma/gypsies. *Am J Hum Genet* [Internet] 75:596–609. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1182047&tool=pmcentrez&rendertype=abstract

Morin SJ, Eccles J, Iturriaga A, Zimmerman RS. 2017. Translocations, inversions and other chromosome rearrangements. *Fertil Steril* 107:19–26.

Morris-Rosendahl DJ, Crocq M-A. 2020. Neurodevelopmental disorders—the history and future of a diagnostic concept. *Dialogues Clin Neurosci* 22:65–72.

Munro AW, Lindsay JG. 1996. Bacterial cytochromes P-450. *Mol Microbiol* 20:1115–1125.

Munsterhjelm M. 2015. Beyond the Line: Violence and the Objectification of the Karitiana Indigenous People as Extreme Other in Forensic Genetics. *International Journal for the Semiotics of Law* [Internet] 28:289–316. Available from: https://link.springer.com/article/10.1007/s11196-014-9395-4

Murakami R. 1987. A histological study of the development of the penis of wild-type and androgen-insensitive mice. *J Anat* 153:223–231.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nat Rev Genet* 14:157–167.

Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S, et al. 2015. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications 2015 6:1* [Internet] 6:1–13. Available from: https://www.nature.com/articles/ncomms9018

Napoli I, Mercaldo V, Boyl PP, Eleuteri B, Zalfa F, de Rubeis S, di Marino D, Mohr E, Massimi M, Falconi M, et al. 2008. The Fragile X Syndrome Protein Represses Activity-Dependent Translation through CYFIP1, a New 4E-BP. *Cell* 134:1042–1054.

Nazaryan-Petersen L, Bjerregaard VA, Nielsen FC, Tommerup N, Tümer Z. 2020. Chromothripsis and DNA Repair Disorders. *J Clin Med* 9:613.

Need AC, Goldstein DB. 2009. Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics* 25:489–494.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet* 11:265–289.

Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP. 2008. Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res* [Internet] 18:1711–1723. Available from: https://genome.cshlp.org/content/18/11/1711.full

Nielsen R. 2004. Population genetic analysis of ascertained SNP data. *Hum Genomics* 1:218.

Niu A, Wang Y, Zhang H, Liao C, Wang J, Zhang R, Che J, Su B. 2011. Rapid evolution and copy number variation of primate RHOXF2, an X-linked homeobox gene involved in male reproduction and possibly brain function. *BMC Evol Biol* 11:298.

Normile D. 2021. Genetic papers containing data from China's ethnic minorities draw fire. *Science (1979)* 373:727–728.

Nothnagel M, Fan G, Guo F, He Y, Hou Y, Hu S, Huang J, Jiang X, Kim W, Kim K, et al. 2022. Retraction Note to: Revisiting the male

genetic landscape of China: a multi-center study of almost 38,000 Y-STR haplotypes(Hum Genet, (2017), 136, (485–497), 10.1007/s00439-017-1759-x). *Hum Genet* [Internet] 141:175–176. Available from: https://link.springer.com/article/10.1007/s00439-021-02413-w

Nowell PC, Hungerford DA. 1960. Chromosome Studies on Normal and Leukemic Human Leukocytes. *JNCI: Journal of the National Cancer Institute*.

Nunes MA, Kučerová K, Lukáč O, Kvapil M, Brož J. 2018. Prevalence of Diabetes Mellitus among Roma Populations—A Systematic Review. *International Journal of Environmental Research and Public Health 2018, Vol. 15, Page 2607* [Internet] 15:2607. Available from: https://www.mdpi.com/1660-4601/15/11/2607/htm

Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Narvaiza I, Sudmant PH, Penn O, Chiatante G, Malig M, Huddleston J, et al. 2016. Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* 536:205–209.

Nyangiri OA, Noyes H, Mulindwa J, Ilboudo H, Kabore JW, Ahouty B, Koffi M, Asina OF, Mumba D, Ofon E, et al. 2020. Copy number variation in human genomes from three major ethno-linguistic groups in Africa. *BMC Genomics* 21:1–15.

Oben B, Froyen G, Maclachlan KH, Leongamornlert D, Abascal F, Zheng-Lin B, Yellapantula V, Derkach A, Geerdens E, Diamond BT, et al. 2021. Whole-genome sequencing reveals progressive versus stable myeloma precursor conditions as two distinct entities. *Nat Commun* 12:1861.

Oguro-Ando A, Rosensweig C, Herman E, Nishimura Y, Werling D, Bill BR, Berg JM, Gao F, Coppola G, Abrahams BS, et al. 2015. Increased CYFIP1 dosage alters cellular and dendritic morphology and dysregulates mTOR. *Mol Psychiatry* 20:1069–1078.

Open Society Foundations. 2015. European Commission Targets Slovakia over Roma School Discrimination. *Open Society Foundations* [Internet]. Available from: https://www.opensocietyfoundations.org/newsroom/european-commission-targets-slovakia-over-roma-school-discrimination

Orr TJ, Brennan PLR. 2016. All Features Great and Small—the Potential Roles of the Baculum and Penile Spines in Mammals. *Integr Comp Biol* 56:635–643.

Ottaviani D, LeCain M, Sheer D. 2014. The role of microhomology in genomic structural variation. *Trends in Genetics* 30:85–94.

Owen MJ, O'Donovan MC. 2017. Schizophrenia and the neurodevelopmental continuum:evidence from genomics. *World Psychiatry* 16:227–235.

Owen MJ, O'Donovan MC, Thapar A, Craddock N. 2011. Neurodevelopmental hypothesis of schizophrenia. *British Journal of Psychiatry* 198:173–175.

Owen MJ, Sawa A, Mortensen PB. 2016. Schizophrenia. *The Lancet* 388:86–97.

Pajic P, Pavlidis P, Dean K, Neznanova L, Romano R-A, Garneau D, Daugherity E, Globig A, Ruhl S, Gokcumen O. 2019. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *Elife* 8.

Pamjav H, Zalán A, Béres J, Nagy M, Chang YM. 2011. Genetic structure of the paternal lineage of the Roma People. *Am J Phys Anthropol* 145:21–29.

Pan X, Liu Changhui, Du W, Chen L, Han X, Yang X, Liu Chao. 2020. RETRACTED ARTICLE: Genetic analysis and forensic evaluation of 47 autosomal InDel markers in four different Chinese populations. *Int J Legal Med* [Internet] 134:2079. Available from: https://link.springer.com/article/10.1007/s00414-019-02059-4

Parekh N, Rose T. 2011. Health Inequalities of the Roma in Europe: a Literature Review. *Cent Eur J Public Health* [Internet] 19:139–142. Available from: http://www.neurope.eu/articles/History-teaches-us-

Parikh H, Mohiyuddin M, Lam HYK, Iyer H, Chen D, Pratt M, Bartha G, Spies N, Losert W, Zook JM, et al. 2016. Svclassify: A method to establish benchmark structural variant calls. *BMC Genomics* [Internet] 17:1–16. Available from: http://dx.doi.org/10.1186/s12864-016-2366-2

Patrinos GP, Pasparakis E, Koiliari E, Pereira AC, Hünemeier T, Pereira L v., Mitropoulou C. 2020. Roadmap for Establishing Large-Scale

Genomic Medicine Initiatives in Low- and Middle-Income Countries. *The American Journal of Human Genetics* 107:589–595.

Pelak K, Need AC, Fellay J, Shianna K v., Feng S, Urban TJ, Ge D, de Luca A, Martinez-Picado J, Wolinsky SM, et al. 2011. Copy Number Variation of KIR Genes Influences HIV-1 Control. *PLoS Biol* 9:e1001208.

Pellestor F. 2019. Chromoanagenesis: cataclysms behind complex chromosomal rearrangements. *Mol Cytogenet* 12:6.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39:1256–1260.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, et al. 2008. Copy number variation and evolution in humans and chimpanzees. *Genome Res* 18:1698–1710.

Petraki I. 2020. Roma Health Mediators: A Neocolonial Tool for the Reinforcement of Epistemic Violence? *Critical Romani Studies* [Internet] 3:72–95. Available from: https://crs.ceu.edu/index.php/crs/article/view/60

Pevzner PA, Compeau GTPEC. 2016. Veritas Genetics. Veritas genetics launches 999 whole genome and sets new standard for genetic testing—Press Release.

Piazza A, Heyer W-D. 2019. Homologous Recombination and the Formation of Complex Genomic Rearrangements. *Trends Cell Biol* 29:135–149.

Piccolo F, Jeanpierre M, Leturcq F, Dodé C, Azibi K, Toutain A, Merlini L, Jarre L, Navarro C, Krishnamoorthy R, et al. 1996. A Founder Mutation in the γ-Sarcoglycan Gene of Gypsies Possibly Predating Their Migration Out of India. *Hum Mol Genet* [Internet] 5:2019–2022. Available from: https://academic.oup.com/hmg/article/5/12/2019/658165

Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J, Gray J. 1988. Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of

chromosome 4. *Proceedings of the National Academy of Sciences* 85:9138–9142.

Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, Thiruvahindrapuram B, Xu X, Ziman R, Wang Z, et al. 2014. Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. *The American Journal of Human Genetics* 94:677–694.

Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, et al. 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466:368–372.

Polley S, Louzada S, Forni D, Sironi M, Balaskas T, Hains DS, Yang F, Hollox EJ. 2015. Evolution of the rapidly mutating human salivary agglutinin gene (DMBT1) and population subsistence strategy. *Proceedings of the National Academy of Sciences* 112:5105–5110.

Poole AC, Goodrich JK, Youngblut ND, Luque GG, Ruaud A, Sutter JL, Waters JL, Shi Q, El-Hadidi M, Johnson LM, et al. 2019. Human Salivary Amylase Gene Copy Number Impacts Oral and Gut Microbiomes. *Cell Host Microbe* 25:553-564.e7.

Popadić A, Anderson WW. 1995. Evidence for gene conversion in the amylase multigene family of Drosophila pseudoobscura. *Mol Biol Evol*.

Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature* 538:161–164.

Portmann T, Yang M, Mao R, Panagiotakos G, Ellegood J, Dolen G, Bader PL, Grueter BA, Goold C, Fisher E, et al. 2014. Behavioral Abnormalities and Circuit Defects in the Basal Ganglia of a Mouse Model of 16p11.2 Deletion Syndrome. *Cell Rep* 7:1077–1092.

Pott AF. 1844. Die Zigeuner In Europa Und Asien: Ethnographisch-linguistische Untersuchung, Vornehmlich Ihrer Herkunft Und Sprache. Heynemann

Prakobphol A, Xu F, Hoang VM, Larsson T, Bergstrom J, Johansson I, Frängsmyr L, Holmskov U, Leffler H, Nilsson C, et al. 2000. Salivary Agglutinin, Which Binds Streptococcus mutansand Helicobacter pylori, Is the Lung Scavenger Receptor Cysteine-rich Protein gp-340. *Journal of Biological Chemistry* 275:39860–39866.

Price G. 2000. Languages in Britain and Ireland. Wiley-Blackwell

Pruimboom L, Fox T, Muskiet FAJ. 2014. Lactase persistence and augmented salivary alpha-amylase gene copy numbers might have been selected by the combined toxic effects of gluten and (food born) pathogens. *Med Hypotheses* 82:326–334.

Qi Y, Martin MP, Gao X, Jacobson L, Goedert JJ, Buchbinder S, Kirk GD, O'Brien SJ, Trowsdale J, Carrington M. 2006. KIR/HLA Pleiotropism: Protection against Both HIV and Opportunistic Infections. *PLoS Pathog* 2:e79.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Rai N, Chaubey G, Tamang R, Pathak AK, Singh VK, Karmin M, Singh M, Rani DS, Anugula S, Yadav BK, et al. 2012. The Phylogeography of Y-Chromosome Haplogroup H1a1a-M82 Reveals the Likely Indian Origin of the European Romani Populations. *PLoS One* 7:1–7.

Ramírez B, Niño-Orrego MJ, Cárdenas D, Ariza KE, Quintero K, Contreras Bravo NC, Tamayo-Agudelo C, González MA, Laissue P, Fonseca Mendoza DJ. 2019. Copy number variation profiling in pharmacogenetics CYP-450 and GST genes in Colombian population. *BMC Med Genomics* [Internet] 12. Available from: https://pubmed.ncbi.nlm.nih.gov/31324178/

Ravnan JB. 2006. Subtelomere FISH analysis of 11 688 cases: an evaluation of the frequency and pattern of subtelomere rearrangements in individuals with developmental disabilities. *J Med Genet* 43:478–489.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* 444:444–454.

Rees E, Walters JTR, Georgieva L, Isles AR, Chambert KD, Richards AL, Mahoney-Davies G, Legge SE, Moran JL, McCarroll SA, et al. 2014. Analysis of copy number variations at 15 schizophrenia-associated loci. *British Journal of Psychiatry* 204:108–114.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature 2009 461:7263*

[Internet] 461:489–494. Available from: https://www.nature.com/articles/nature08365

Reich DE, Cargili M, Boik S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, et al. 2001. Linkage disequilibrium in the human genome. *Nature 2001 411:6834* [Internet] 411:199–204. Available from: https://www.nature.com/articles/35075590

Reichhardt MP, Holmskov U, Meri S. 2017. SALSA—A dance on a slippery floor with changing partners. *Mol Immunol* 89:100–110.

Reiter LT, Hastings PJ, Nelis E, de Jonghe P, van Broeckhoven C, Lupski JR. 1998. Human Meiotic Recombination Products Revealed by Sequencing a Hotspot for Homologous Strand Exchange in Multiple HNPP Deletion Patients. *The American Journal of Human Genetics* 62:1023–1033.

Reiter LT, Murakami T, Koeuth T, Pentao L, Muzny DM, Gibbs RA, Lupski JR. 1996. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nat Genet* 12:288–297.

Relethford JH, Harding RM. 2001. Population Genetics of Modern Human Evolution. *eLS* [Internet]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1038/npg.els.0001470

Rex-Kiss B, Szabo L, Szabo S, Hartmann É. 1973. ABO, MN, Rh blood groups, Hp types and Hp level, Gm (1) factor investigations on the Gypsy population of Hungary. *Hum Biol*:41–61.

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 13:278–289.

Robyt JF, French D. 1967. Multiple attack hypothesis of α-amylase action: Action of porcine pancreatic, human salivary, and Aspergillus oryzae α-amylases. *Arch Biochem Biophys* 122:8–16.

Ronald A, Hoekstra RA. 2011. Autism spectrum disorders and autistic traits: A decade of new twin studies. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 156:255–274.

Rowley JD. 1973. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature* 243:290–293.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko J v., Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature 2002 419:6909* [Internet] 419:832–837. Available from: https://www.nature.com/articles/nature01140

Saitou M, Gokcumen O. 2019a. An Evolutionary Perspective on the Impact of Genomic Copy Number Variation on Human Health. *J Mol Evol* [Internet]. Available from: https://doi.org/10.1007/s00239-019-09911-6

Saitou M, Gokcumen O. 2019b. Resolving the Insertion Sites of Polymorphic Duplications Reveals a HERC2 Haplotype under Selection. *Genome Biol Evol* [Internet] 11:1679–1690. Available from: https://academic.oup.com/gbe/article/11/6/1679/5498151

Sampson J. 1923. On the origin and early migrations of the Gypsies. *Romani Studies* 2:156.

Sampson J. 1926. The dialect of the Gypsies of Wales. Oxford: Clarendon

Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, Shamardina O, Stirrups K, Delon I, Dewhurst E, et al. 2018. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med* [Internet] 10:1–10. Available from: https://link.springer.com/articles/10.1186/s13073-018-0606-6

Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, et al. 2015. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 87:1215–1233.

Sasaki H, Betensky RA, Cairncross JG, Louis DN. 2002. DMBT1 Polymorphisms: Relationship to Malignant Glioma Tumorigenesis1. *Cancer Res* 62:1790–1796.

Sasaki T, Rodig SJ, Chirieac LR, Jänne PA. 2010. The biology and treatment of EML4-ALK non-small cell lung cancer. *Eur J Cancer* 46:1773–1780.

Scelza BA, Atkinson EG, Prall S, McElreath R, Sheehama J, Henn BM. 2020. The ethics and logistics of field-based genetic paternity studies. *Evol Hum Sci* [Internet] 2:e22. Available from: https://www.cambridge.org/core/journals/evolutionary-human-

sciences/article/ethics-and-logistics-of-fieldbased-genetic-paternity-studies/1D20DEC299596C012301CF9AC949970D

Scherer A. 2009. Batch Effects and Noise in Microarray Experiments: Sources and Solutions. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*:1–260.

Schröder J, Hsu A, Boyle SE, Macintyre G, Cmero M, Tothill RW, Johnstone RW, Shackleton M, Papenfuss AT. 2014. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics* [Internet] 30:1064–1072. Available from: https://academic.oup.com/bioinformatics/article/30/8/1064/257640

Scott SA, Scott ER, Seki Y, Chen AJ, Wallsten R, Owusu Obeng A, Botton MR, Cody N, Shi H, Zhao G, et al. 2021. Development and Analytical Validation of a 29 Gene Clinical Pharmacogenetic Genotyping Panel: Multi-Ethnic Allele and Copy Number Variant Detection. *Clin Transl Sci* [Internet] 14:204–213. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/cts.12844

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong Association of De Novo Copy Number Mutations with Autism. *Science (1979)* 316:445–449.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. 2004. Large-Scale Copy Number Polymorphism in the Human Genome. *Science (1979)* 305:525–528.

Sebat J, Levy DL, McCarthy SE. 2009. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends in Genetics* 25:528–535.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* 19:329–346.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018a. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods 2018 15:6* [Internet] 15:461–468. Available from: https://www.nature.com/articles/s41592-018-0001-7

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018b. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 15:461–468.

Sekiguchi M, Sobue A, Kushima I, Wang C, Arioka Y, Kato H, Kodama A, Kubo H, Ito N, Sawahata M, et al. 2020. ARHGAP10, which encodes Rho GTPase-activating protein 10, is a novel gene for schizophrenia risk. *Transl Psychiatry* 10:247.

Seo JS, Rhie A, Kim Junsoo, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim Jihye, et al. 2016. De novo assembly and phasing of a Korean human genome. *Nature 2016 538:7624* [Internet] 538:243–247. Available from: https://www.nature.com/articles/nature20098

Shaikh TH, Kurahashi H, Saitta SC, O'Hare AM, Hu P, Roe BA, Driscoll DA, McDonald-McGinn DM, Zackai EH, Budarf ML, et al. 2000. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* [Internet] 9:489–501. Available from: https://academic.oup.com/hmg/article/9/4/489/2901343

Shao L, Shaw CA, Lu X-Y, Sahoo T, Bacino CA, Lalani SR, Stankiewicz P, Yatsenko SA, Li Y, Neill S, et al. 2008. Identification of chromosome abnormalities in subtelomeric regions by microarray analysis: A study of 5,380 cases. *Am J Med Genet A* 146A:2242–2251.

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, Schroer RJ, Novara F, de Gregori M, Ciccone R, et al. 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* 40:322–328.

Shaw CJ. 2004. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet* 13:57R – 64.

Shen MM. 2013a. Chromoplexy: A New Category of Complex Rearrangements in the Cancer Genome. *Cancer Cell* 23:567–569.

Shen MM. 2013b. Chromoplexy: A New Category of Complex Rearrangements in the Cancer Genome. *Cancer Cell* 23:567–569.

Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, et al. 2016. Long-read sequencing and de novo assembly of

a Chinese genome. *Nature Communications 2016 7:1* [Internet] 7:1–10. Available from: https://www.nature.com/articles/ncomms12065

Shinawi M, Schaaf CP, Bhatt SS, Xia Z, Patel A, Cheung SW, Lanpher B, Nagl S, Herding HS, Nevinny-Stickel C, et al. 2009. A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes. *Nat Genet* 41:1269–1271.

Simko V, Ginter E. 2010. Short Life Expectancy and Metabolic Syndrome in Romanies (Gypsies) in Slovakia. *Cent Eur J Public Health* [Internet] 18:16–18. Available from: http://www.euro.who.int/HFADB.

Singh T, Walters JTR, Johnstone M, Curtis D, Suvisaari J, Torniainen M, Rees E, Iyegbe C, Blackwood D, McIntosh AM, et al. 2017. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat Genet* 49:1167–1173.

Sirugo G, Williams SM, Tishkoff SA. 2019. The Missing Diversity in Human Genetic Studies. *Cell* [Internet] 177:26–31. Available from: https://doi.org/10.1016/j.cell.2019.02.048

Sivakova D. 1983. Distribution of three red-cell enzyme polymorphisms (ACP, PGM1 and AK) in gypsies from Slovakia (Czechoslovakia). *Ann Hum Biol* 10:449–452.

Smajlagić D, Lavrichenko K, Berland S, Helgeland Ø, Knudsen GP, Vaudel M, Haavik J, Knappskog PM, Njølstad PR, Houge G, et al. 2021. Population prevalence and inheritance pattern of recurrent CNVs associated with neurodevelopmental disorders in 12,252 newborns and their parents. *European Journal of Human Genetics* 29:205–215.

Smigielski EM, Sirotkin K, Ward M, Sherry ST. 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* [Internet] 28:352–355. Available from: https://academic.oup.com/nar/article/28/1/352/2384425

Sobue A, Kushima I, Nagai T, Shan W, Kohno T, Aleksic B, Aoyama Y, Mori D, Arioka Y, Kawano N, et al. 2018. Genetic and animal model analyses reveal the pathogenic role of a novel deletion of RELN in schizophrenia. *Sci Rep* 8:13046.

Spielmann M, Lupiáñez DG, Mundlos S. 2018. Structural variation in the 3D genome. *Nature Reviews Genetics 2018 19:7* [Internet] 19:453–

467. Available from: https://www.nature.com/articles/s41576-018-0007-0

Sridhar CR. 2006. Historical Amnesia: The Romani Holocaust. *Econ Polit Wkly* [Internet] 41:3569–3571. Available from: http://www.jstor.org.proxy.library.georgetown.edu/stable/4418585

Stallings RL, Whitmore SA, Doggett NA, Callen DF. 1993. Refined physical mapping of chromosome 16-specific low-abundance repetitive DNA sequences. *Cytogenet Genome Res* 63:97–101.

Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* 18:74–82.

Stankiewicz P, Lupski JR. 2010. Structural Variation in the Human Genome and its Role in Disease. *Annu Rev Med* 61:437–455.

Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, Bjornsdottir G, Walters GB, Jonsdottir GA, Doyle OM, et al. 2014. CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* 505:361–366.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al. 2011. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* 144:27–40.

Storlazzi CT, Lonoce A, Guastadisegni MC, Trombetta D, D'Addabbo P, Daniele G, L'Abbate A, Macchia G, Surace C, Kok K, et al. 2010. Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res* 20:1198–1206.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazlsy C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene phenotypes. *Science (1979)* 315:848–853.

Sturtevant AH. 1921. A Case of Rearrangement of Genes in Drosophila. *Proceedings of the National Academy of Sciences* 7:235–237.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. 2010. Diversity of human copy number variation and multicopy genes. *Science (1979)* 330:641–646.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science (1979)* 349.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81.

Sullivan PF, Kendler KS, Neale MC. 2003. Schizophrenia as a Complex Trait. *Arch Gen Psychiatry* 60:1187.

Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. *Hum Mol Genet* [Internet] 10:591–597. Available from: https://academic.oup.com/hmg/article/10/6/591/628411

Sunyaev SR, Lathe WC, Ramensky VE, Bork P. 2000. SNP frequencies in human genes: An excess of rare alleles and differing modes of selection. *Trends in Genetics* [Internet] 16:335–337. Available from: http://www.cell.com/article/S0168952500020588/fulltext

Swaminathan GJ, Bragin E, Chatzimichali EA, Corpas M, Bevan AP, Wright CF, Carter NP, Hurles ME, Firth H v. 2012. DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Hum Mol Genet* [Internet] 21:R37–R44. Available from: https://academic.oup.com/hmg/article/21/R1/R37/658334

Takumi T. 2011. The neurobiology of mouse models syntenic to human chromosome 15q. *J Neurodev Disord* 3:270–281.

Takumi T, Tamada K. 2018. CNV biology in neurodevelopmental disorders. *Curr Opin Neurobiol* 48:183–192.

Tansey KE, Rees E, Linden DE, Ripke S, Chambert KD, Moran JL, McCarroll SA, Holmans P, Kirov G, Walters J, et al. 2016. Common alleles contribute to schizophrenia in CNV carriers. *Mol Psychiatry* 21:1085–1089.

Tarkhnishvili D, Gavashelishvili A, Murtskhvaladze M, Gabelaia M, Tevzadze G. 2014. Human paternal lineages, languages, and environment in the Caucasus. *Hum Biol* 86:113–130.

Tattini L, D'Aurizio R, Magi A. 2015. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol* [Internet] 3:1–8. Available from: http://journal.frontiersin.org/Article/10.3389/fbioe.2015.00092/abstract

Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM, et al. 2010. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences* 107:10848–10853.

Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C, et al. 2016. Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A* [Internet] 113:11901–11906. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1613365113

Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* [Internet] 28:2711–2718. Available from: https://academic.oup.com/bioinformatics/article/28/21/2711/237315

The International Schizophrenia Consortium. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* [Internet] 455:237–241. Available from: http://www.nature.com/articles/nature07239

Tiffin N. 2018. Tiered informed consent: respecting autonomy, agency and individuality in Africa. *BMJ Glob Health* [Internet] 3:e001249. Available from: https://gh.bmj.com/content/3/6/e001249

Trask BJ. 2002. Human cytogenetics: 46 chromosomes, 46 years and counting. *Nature Reviews Genetics 2002 3:10* [Internet] 3:769–778. Available from: https://www.nature.com/articles/nrg905

Tsosie KS, Yracheta JM, Kolopenuk JA, Geary J. 2021. We Have "Gifted" Enough: Indigenous Genomic Data Sovereignty in Precision Medicine. *https://doi.org/10.1080/15265161.2021.1891347* [Internet] 21:72–75. Available from: https://www.tandfonline.com/doi/abs/10.1080/15265161.2021.1891347

Turajlic S, Sottoriva A, Graham T, Swanton C. 2019. Resolving genetic heterogeneity in cancer. *Nat Rev Genet* 20:404–416.

Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, Beck S, Hurles ME. 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* 40:90–95.

Turner RL. 1926. The position of Romani in Indo-Aryan. *Journal of the Gypsy Lore Society. Third series* V:145–189.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* 37:727–732.

Tyson C, Harvard C, Locker R, Friedman JM, Langlois S, Lewis MES, van Allen M, Somerville M, Arbour L, Clarke L, et al. 2005. Submicroscopic deletions and duplications in individuals with intellectual disability detected by array-CGH. *Am J Med Genet A* 139A:173–185.

Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, Cao H, Moon JE, Kashin S, Fuchsberger C, et al. 2015. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat Genet* 47:921–925.

Vendramin R, Litchfield K, Swanton C. 2021. Cancer evolution: Darwin and beyond. *EMBO J* 40.

Verhaak RGW, Bafna V, Mischel PS. 2019. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat Rev Cancer* 19:283–288.

Vozarova De Courten B, de Courten M, Hanson RL, Zahorakova A, Egyenes HP, Tataranni PA, Bennett PH, Vozar J. 2003. Higher prevalence of type 2 diabetes, metabolic syndrome and cardiovascular diseases in gypsies than in non-gypsies in Slovakia. *Diabetes Res Clin Pract* 62:95–103.

Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K. 2001. The discovery of single-nucleotide polymorphisms—And inferences about human demographic history. *Am J Hum Genet* [Internet] 69:1332–1347. Available from: http://www.cell.com/article/S0002929707612622/fulltext

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M. 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* [Internet] 17:1665–1674. Available from: https://genome.cshlp.org/content/17/11/1665.full

Weckselblatt B, Hermetz KE, Rudd MK. 2015. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res* 25:937–947.

Weckselblatt B, Rudd MK. 2015a. Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends in Genetics* 31:587–599.

Weckselblatt B, Rudd MK. 2015b. Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends in Genetics* 31:587–599.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat Rev Genet* [Internet] 14:125–138. Available from: http://dx.doi.org/10.1038/nrg3373

Weise A, Mrasek K, Klein E, Mulatinho M, Llerena JC, Hardekopf D, Pekova S, Bhatt S, Kosyakova N, Liehr T. 2012. Microdeletion and Microduplication Syndromes. *Journal of Histochemistry & Cytochemistry* 60:346–358.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MAR, Green T, et al. 2008. Association between Microdeletion and Microduplication at 16p11.2 and Autism. *New England Journal of Medicine* 358:667–675.

Wells RD. 2007. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem Sci* 32:271–278.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 37:1155–1162.

Werissa NA, Piko P, Fiatal S, Kosa Z, Sandor J, Adany R. 2019. SNP-Based Genetic Risk Score Modeling Suggests No Increased Genetic

Susceptibility of the Roma Population to Type 2 Diabetes Mellitus. *Genes 2019, Vol. 10, Page 942* [Internet] 10:942. Available from: https://www.mdpi.com/2073-4425/10/11/942/htm

Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer RM, Markenscoff-Papadimitriou E, et al. 2018. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* [Internet] 50:727–736. Available from: http://dx.doi.org/10.1038/s41588-018-0107-y

Wetterstrand KA. 2013. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP).

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–876.

Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, Charlson FJ, Norman RE, Flaxman AD, Johns N, et al. 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet* 382:1575–1586.

Williams F, Maxwell LD, Halfpenny IA, Meenagh A, Sleator C, Curran MD, Middleton D. 2003. Multiple copies of KIR 3DL/S1 and KIR 2DL4 genes identified in a number of individuals. *Hum Immunol* 64:729–732.

Williams TN, Wambua S, Uyoga S, Macharia A, Mwacharo JK, Newton CRJC, Maitland K. 2005. Both heterozygous and homozygous α+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. *Blood* [Internet] 106:368–371. Available from: https://ashpublications.org/blood/article/106/1/368/103158/Both-heterozygous-and-homozygous-thalassemias

Wong K, Keane TM, Stalker J, Adams DJ. 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol* [Internet] 11:1–9. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-12-r128

Wu S, Turner KM, Nguyen N, Raviram R, Erb M, Santini J, Luebeck J, Rajkumar U, Diao Y, Li B, et al. 2019. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* 575:699–703.

Xiang Y, Niu Y, Xie Y, Chen S, Zhu F, Shen W, Zeng L. 2021. Inhibition of RhoA/Rho kinase signaling pathway by fasudil protects against kainic acid-induced neurite injury. *Brain Behav* 11.

Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, Huang N, Zerjal T, Lee C, Carter NP, et al. 2008. Adaptive Evolution of UGT2B17 Copy-Number Variation. *The American Journal of Human Genetics* 83:337–346.

Yao Y, Wang H, Li B, Tang Y. 2014. Evaluation of the TMPRSS2:ERG fusion for the detection of prostate cancer: a systematic review and meta-analysis. *Tumor Biology* 35:2157–2166.

Yasukochi Y, Satta Y. 2011. Evolution of the CYP2D gene cluster in humans and four non-human primates. *Genes Genet Syst* 86:109–116.

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* 19:1586–1592.

Young JM, Endicott RLM, Parghi SS, Walker M, Kidd JM, Trask BJ. 2008. Extensive Copy-Number Variation of the Human Olfactory Receptor Gene Family. *The American Journal of Human Genetics* 83:228–242.

Zalán A, Béres J, Pamjav H. 2011. Paternal genetic history of the Vlax Roma. *Forensic Sci Int Genet* 5:109–113.

Zanger UM, Schwab M. 2013. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther* 138:103–141.

Zankl H, Zang KD. 1971. Structural variability of the normal human karyotype. *Humangenetik* 13:160–162.

Zeggini E. 2014. Using genetically isolated populations to understand the genomic basis of disease. *Genome Med* 6:12–14.

Zeljko H, Kari-Juri T, Smolej Narani N, Salihovi MP, Klari IM, Barbali M, Starevi B, Lauc LB, Janiijevi B. 2008. Traditional CVD Risk

Factors and Socio-Economic Deprivation in Roma Minority Population of Croatia. *Coll. Antropol* 32:667–676.

Zepeda-Mendoza CJ, Morton CC. 2019. The Iceberg under Water: Unexplored Complexity of Chromoanagenesis in Congenital Disorders. *The American Journal of Human Genetics* 104:565–577.

Zhai Y, Zhang Z, Shi P, Martin DM, Kong X. 2021. Incorporation of exome-based CNV analysis makes trio-WES a more powerful tool for clinical diagnosis in neurodevelopmental disorders: A retrospective study. *Hum Mutat* 42:990–1004.

Zhang D, Cheng L, Qian Y, Alliey-Rodriguez N, Kelsoe JR, Greenwood T, Nievergelt C, Barrett TB, McKinney R, Schork N, et al. 2009. Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol Psychiatry* 14:376–380.

Zhang H, Jain C, Aluru S. 2020. A comprehensive evaluation of long read error correction methods. *BMC Genomics* [Internet] 21:1–15. Available from: https://link.springer.com/articles/10.1186/s12864-020-07227-0

Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW. 2006. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet Genome Res* [Internet] 115:205–214. Available from: https://www.karger.com/Article/FullText/95916

Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, Niu Y, Shi Y, Zhou H, Song T, et al. 2021. NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep* [Internet] 37:110017. Available from: http://www.cell.com/article/S2211124721014996/fulltext

Zhang X, Sun H, Danila DC, Johnson SR, Zhou Y, Swearingen B, Klibanski A. 2002. Loss of Expression of GADD45γ , a Growth Inhibitory Gene, in Human Pituitary Adenomas: Implications for Tumorigenesis. *J Clin Endocrinol Metab* 87:1262–1267.

Zhao M, Wang Qingguo, Wang Quan, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics* [Internet] 14:S1. Available from: http://www.biomedcentral.com/1471-2105/14/S11/S1

Zhao X, Collins RL, Lee W-P, Weber AM, Jun Y, Zhu Q, Weisburd B, Huang Y, Audano PA, Wang H, et al. 2021. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *The American Journal of Human Genetics* 108:919–928.

Zhao X, Li C, Paez JG, Chin K, Jänne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, et al. 2004. An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Res* [Internet] 64:3060–3071. Available from: https://aacrjournals.org/cancerres/article/64/9/3060/517887/An-Integrated-View-of-Copy-Number-and-Allelic

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology 2016 34:3* [Internet] 34:303–311. Available from: https://www.nature.com/articles/nbt.3432

Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* [Internet] 38:1347–1355. Available from: http://dx.doi.org/10.1038/s41587-020-0538-8

7. APPENDIX

7.1 Supplementary methods

Structural variants calling algorithms

Different methods exist to call structural variants (SVs) from short read sequencing data. To infer the presence of structural sequence changes in a sample (or a group of samples), a number of methodologies leverage in different ways the information from mapped sequences. The most frequently used algorithms devised so far are based on read depth, split reads, read pairs or assembly approaches. Briefly, read-depth method assumes that the coverage of a region relates to the number of copies of that region. To assess this, the method counts the number of reads mapping to fixed size regions (bins) and, after data normalization, estimates the number of copies in the genome (Tattini et al. 2015). The split-read approach takes advantage of how paired-end reads map: if one of the paired reads accurately maps to the reference while the other does not or only maps partially, the latter may signal the presence of an SV breakpoint. The splitting of the unmapped (or partially mapped) read in sub-reads allows for a second mapping step of these portions independently. After this step, the two portions of the split read will flank respectively the start and end point of the detected SV (Zhao et al. 2013). In the read-pair approach, SVs are detected by relying on the spacing between read pairs mapped to reference. Pairs of reads mapping closer or further to one another than what is expected based on their average insert size signal the presence of an SV (Zhao et al. 2013). The last method, assembly-based (or *de novo* assembly), uses groups of overlapping reads to create contigs; these sequences are longer than short reads and represent the union of non-repetitive information that the latter provide. The comparisons between contigs and a reference genome highlight regions with putative discordant copy number, where SV may occur (Tattini et al. 2015). Here we provide the list of software used to call SVs and the settings we used.

CNVnator

We ran CNVnator (version 0.4.1) (Abyzov et al. 2011b) using a bin size of 100bp, identifying duplication and deletion calls for each sample. Results then underwent copy number estimation using the –genotype command and the results were filtered for calls having e-val1, e-val2 and q0 parameters with values lower than 0.05, 0.005 and 0.5 respectively.

BreakDancer

As a first step using the BreakDancer (version 1.4.5) pipeline (Chen et al. 2009), we used the built-in script bam2cfg.pl to obtain relevant data from the input BAM files, such as read length, average insert size and standard deviation, to be used subsequently by the algorithm. We then checked for a number of parameters (i.e., RG or LB information in the header, coefficient of variation of the insert size, percentage of inter-chromosomal read pairs) to assess the quality of the input files. BreakDancer with default parameters generates raw results, subsequently filtered to retain copy numbers equals to zero, one, three or more. For further refinement of the results, we implemented the Perl BreakDown software (version 1.1.1) (Fan et al. 2014) to work with the BreakDancer output, obtaining genotype calls and filtered results. We turned on GC correction (-g option), and filtered for variant score lower than 40, for mapping quality lower than 30 (-q option) and for event size lower than 100bp.

Pindel

Pindel (version 0.2.5b8) (Ye et al. 2009) uses the split-read method, and we ran it with default parameters, including BreakDancer result calls as a support to increase its sensitivity and specificity. Once the main script was executed, we applied filters removing those calls shorter than 100bp, or having a mapping quality lower than 30 and a number of supporting reads

lower than 20% the average coverage of the samples included in the analysis.

Tardis

We first used mrsFAST (version 3.4.0) (Hach et al. 2014) to extract discordant read pairs from the paired-end reads (fastq) files of each sample, providing the read-pair information for subsequent analysis. Discordant reads refer to read pairs spacing outside the minimum-maximum range of the fragment size and mapping respectively to forward and reverse strand. We ran Tardis (version 1.0.4) (Soylev et al. 2017) with default parameters and we excluded variants below 100bp.

Lumpy

To prepare support data for Lumpy (version 0.2.13) (Layer et al. 2014), we used SpeedSeq (version 0.1.2) (Chiang et al. 2015), which takes fastq files as input to generate splitters and discordant files reporting information about split-read and read-pair methods. Lumpy also integrates CNVnator results for further read-depth information. We ran the software with default parameters and excluding low complexity regions of the genome that may lead to unreliable results. The output underwent the SVTyper (version 0.7.1) (Chiang et al. 2015) algorithm to produce genotyped results. We filtered out calls with mapping quality lower than 30, a number of supporting reads lower than 5 and size lower than 100bp.

GenomeSTRiP

We used the GenomeSTRiP (version 2.0) (Handsaker et al. 2011; Handsaker et al. 2015b), which implements the CNVDsicovery pipeline. We first performed data pre-processing with the SVPreprocess pipeline, which generates metadata needed for all the subsequent steps. We then applied the CNVDsicovery pipeline, to call for CNVs in our sample set and

finally run SVGenotyper pipeline to obtain genotyped calls for our results. We finally removed those variants having low quality scores (CNQ > 12; "LQ") from further analyses and events smaller than 100bp.

Batch effect filtering

When using PCA on copy numbers in deletions and duplications, our dataset showed batch effects, clustering the samples based on the dataset they belong to. We attempted to overcome batch effects by refining our genotype calls for deletions and duplications and by subsequently filtering the resulting set of variants. We specifically re-estimated each CNV's genotype for all the individuals in our dataset by using GraphTyper2 (version 2.5.1) (Eggertsson et al. 2019) and, after quality filtering, accurately recovered 7257 variants (6070 deletions and 1187 duplications), filtering out ~35% of the variants present in the initial set. Since a portion of the variants detected were false positives (i.e. initially detected as heterozygous/homozygous for a CNV and subsequently genotyped as homozygous for the reference allele), a deeper scan for such calls was conducted using the R packages CNVfilteR (version 1.8.0) (Moreno-Cabrera et al. 2021) and HardyWeinberg (version 1.7.2) (Graffelman 2015). The former uses SNP data to identify incorrectly called deletions and duplications, while the latter performs chi-squared test for Hardy-Weinberg equilibrium. We filtered out 778 (520 deletions and 258 duplications) and 2819 (2379 deletions and 440 duplications) variants using CNVfilteR and HardyWeinberg packages respectively. The final filtered dataset comprised 3660 CNVs (3171 deletions and 489 duplications). The larger number of deletions compared to duplications probably reflects the higher performance of software to resolve deletion events compared to duplications (Zook et al. 2020; Khayat et al. 2021).

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: An Approach to Discover, Genotype and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing. :974–984.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* [Internet] 6:677–681. Available from: http://dx.doi.org/10.1038/nmeth.1363

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat. Methods* 12:966–968.

Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, Gudbjartsson DF, Stefansson K, Halldorsson B V., Melsted P. 2019. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* [Internet] 10:1–8. Available from: http://dx.doi.org/10.1038/s41467-019-13341-9

Fan X, Zhou W, Chong Z, Nakhleh L, Chen K. 2014. Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinformatics* 15:1–12.

Graffelman J. 2015. Exploring diallelic genetic markers: The HardyWeinberg package. *J. Stat. Softw.* [Internet] 64:1–23. Available from: http://hdl.handle.net/2117/76647

Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2014. MrsFAST-Ultra: A compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res.* 42:494–500.

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, Mccarroll SA. 2015. Large multiallelic copy number variations in humans. *Nat. Genet.* [Internet] 47:296–303. Available from: http://dx.doi.org/10.1038/ng.3200

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* 43:269–276.

Khayat MM, Mohammad S, Sahraeian E, Zarate S, Carroll A, Hong H,

Pan B, Shi L, Gibbs RA, Mohiyuddin M, et al. 2021. Hidden biases in germline structural variant detection. *Genome Biol.* 22:347.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* 15:1–19.

Moreno-Cabrera JM, del Valle J, Castellanos E, Feliubadaló L, Pineda M, Serra E, Capellá G, Lázaro C, Gel B. 2021. CNVfilteR: an R/Bioconductor package to identify false positives produced by germline NGS CNV detection tools. *Bioinformatics*:1–3.

Soylev A, Kockan C, Hormozdiari F, Alkan C. 2017. Toolkit for automated and rapid discovery of structural variants. *Methods* [Internet] 129:3–7. Available from: https://doi.org/10.1016/j.ymeth.2017.05.030

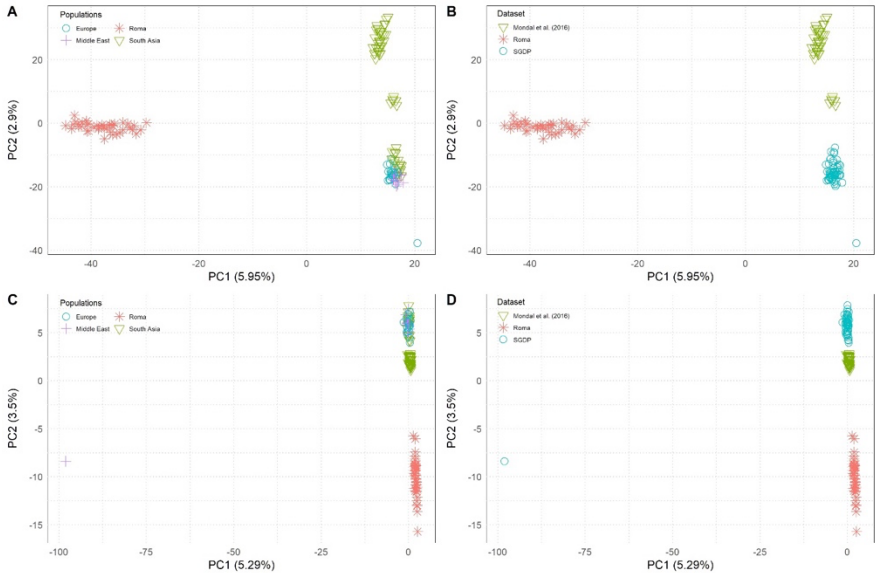
Tattini L, D'Aurizio R, Magi A. 2015. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front. Bioeng. Biotechnol.* [Internet] 3:1–8. Available from: http://journal.frontiersin.org/Article/10.3389/fbioe.2015.00092/abstract

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25:2865–2871.

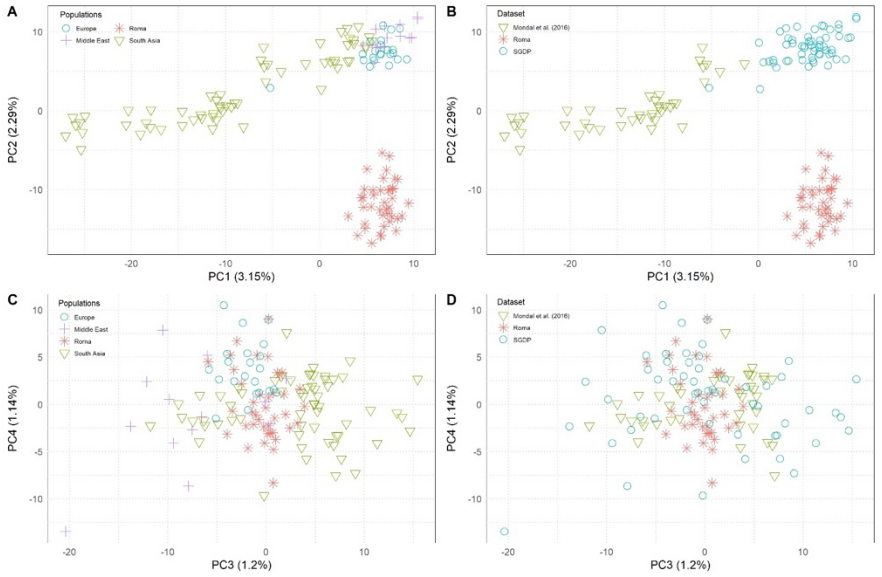
Zhao M, Wang Qingguo, Wang Quan, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics* [Internet] 14:S1. Available from: http://www.biomedcentral.com/1471-2105/14/S11/S1

Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* [Internet] 38:1347–1355. Available from: http://dx.doi.org/10.1038/s41587-020-0538-8

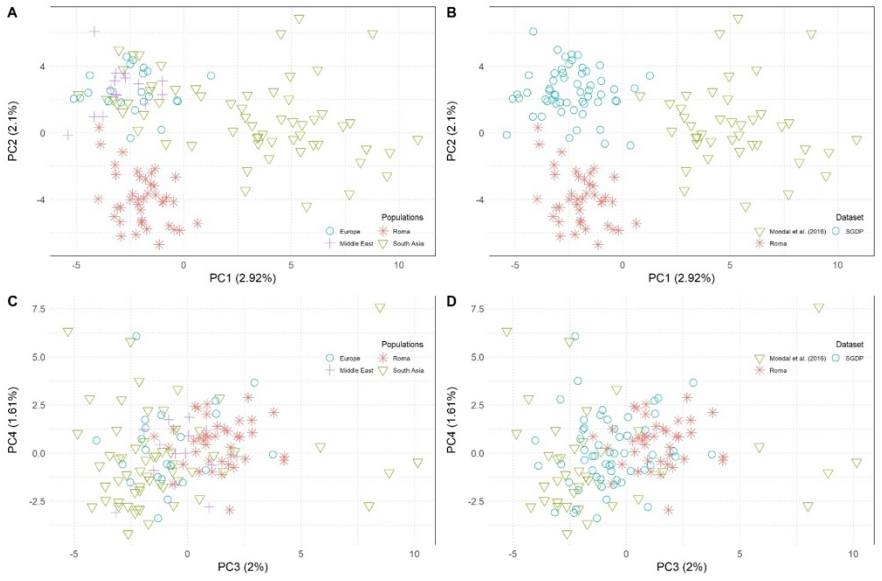
7.2 Supplementary figures



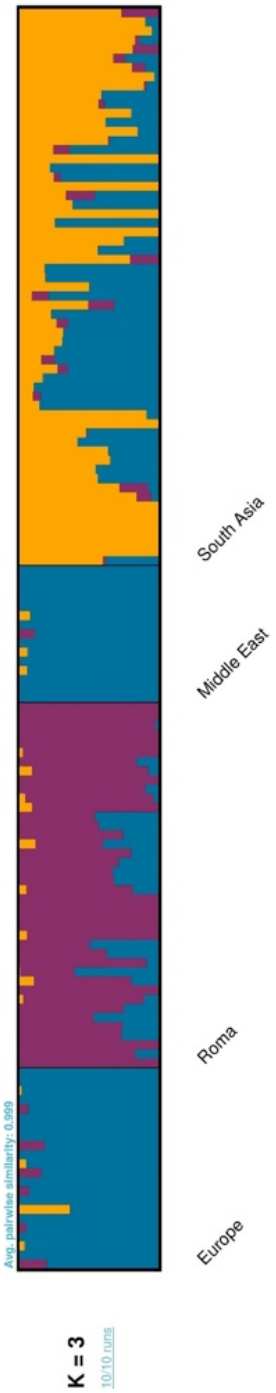
Supplementary Figure 1. PCA of unfiltered dataset. Batch effect in the dataset. PCA plots at the top (A and B) show analysis with deletions, bottom plots (C and D) show duplications. Points shape and colour follow population (A and C) and dataset (B and D) labels



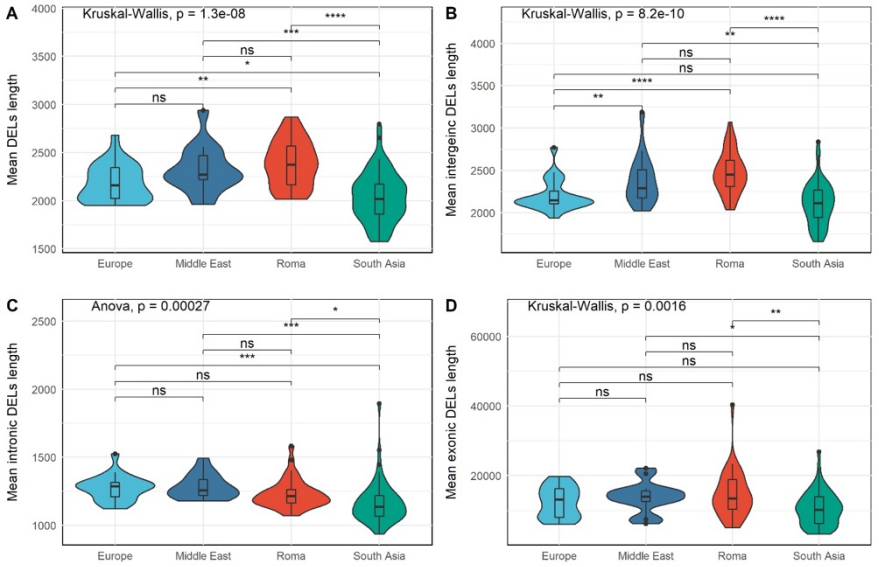
Supplementary Figure 2. PCA of regenotyped and filtered dataset's deletions. PCA plots with population (A, C) and dataset (B, D) labels. Upper plots show principal components 1 and 2, while lower plots show principal components 3 and 4.



Supplementary Figure 3. PCA of regenotyped and filtered dataset's duplications. PCA plots with population (A, C) and dataset (B, D) labels. Upper plots show principal components 1 and 2, while lower plots show principal components 3 and 4.



Supplementary Figure 4. ADMIXTURE analysis. ADMIXTURE plot using K = 3 ancestral populations. Blue, yellow and purple respectively represent West Eurasian, South Asian and Romani ancestries.



Supplementary Figure 5. Mean deletions length distributions among populations. Plots show mean deletions length per individual among populations, p-values for ANOVA or Kruskal-Wallis tests and pairwise post-hoc comparisons. The analyses considered all deletions together (A) and intergenic (B), intronic (C) and exonic (D) deletions.

7.3 Supplementary tables

	Mean (median) Vst	
Pair	**Deletions**	**Duplications**
Roma - Europe	0.039 (0.015)	0.037 (0.011)
Roma - Middle East	0.04 (0.014)	0.036 (0.013)
Roma - South Asia	0.037 (0.014)	0.032 (0.011)
Europe - Middle East	0.03 (0.013)	0.026 (0.011)
Europe - South Asia	0.028 (0.013)	0.024 (0.008)
Middle East - South Asia	0.028 (0.013)	0.026 (0.014)

Supplementary Table 1. Mean and median Vst values of deletion and duplication shared between pairwise populations

Pair	Mean (median) deletions Vst			Mean (median) duplications Vst		
	Intergenic	Intronic	Exonic	Intergenic	Intronic	Exonic
Roma - Europe	0.138 (0.099)	0.132 (0.096)	0.134 (0.116)	0.153 (0.109)	0.122 (0.089)	0.112 (0.07)
Roma - Middle East	0.14 (0.106)	0.145 (0.118)	0.125 (0.102)	0.123 (0.091)	0.123 (0.098)	0.106 (0.082)
Roma - South Asia	0.125 (0.096)	0.131 (0.101)	0.14 (0.131)	0.117 (0.101)	0.083 (0.079)	0.143 (0.146)
Europe - Middle East	0.096 (0.08)	0.095 (0.081)	0.104 (0.103)	0.084 (0.077)	0.081 (0.068)	0.072 (0.072)
Europe - South Asia	0.092 (0.078)	0.089 (0.073)	0.109 (0.089)	0.091 (0.069)	0.065 (0.058)	0.069 (0.068)
Middle East - South Asia	0.089 (0.072)	0.092 (0.072)	0.14 (0.072)	0.087 (0.072)	0.078 (0.07)	0.074 (0.072)

Supplementary Table 2. Top 20% Vst values. Mean and median Vst values of deletion and duplication shared between pairwise populations, divided for genomic localization: intergenic, intronic, exonic

Coordinates	CNV type	Length (kb)	Samples	Genes	TAD coordinates	fRoma	fEurope	fMiddle East	fSouth Asia
1:16550919-16681849	Duplication	130,9	78	CROCCP2, MIR3675, MST1P2, NBPF1, RNU1-1, RNU1-2, RNU1-3, RNU1-4, RNVU1-18	1:16553506-16993505	0,363	0,318	0,300	0,402
1:79159512-79342718	Deletion	183,2	3	ADGRL4	1:79294316-80254315	0,025	0	0,033	0
2:158790463-159097021	Duplication	306,6	1	DAPL1, TANC1	2:158823489-161303489	0	0,023	0	0
4:69243750-69366611	Deletion	122,9	46	UGT2B28	4:69254283-70614283	0,188	0,136	0,167	0,238
7:16097998-16463135	Deletion	365,1	1	CRPPA, CRPPA-AS1, SOSTDC1	7:16120376-16760375	0,013	0	0	0
8:101432838-101573544	Duplication	140,7	3	GRHL2	8:101467773-102267772	0	0	0	0,025
14:105616480-105768229	Deletion	151,7	52	ELK2AP, MIR8071-1, MIR8071-2	14:105653664-106053818	0,313	0,205	0,167	0,107
22:22698419-22893780	Deletion	195,4	2	IGLL5, MIR5571, MIR650	22:22777504-23857813	0	0	0,033	0,008

Supplementary Table 3. Information for eight CNVs intersecting TADs. The table reports CNVs and TADs coordinates, CNV type, length, number of samples having the CNV, the genes intersected by each CNV and its frequency

Coordinates	Length	Tag SNPs	Genes CNVs	Genes SNPs	Common genes	GWAS traits	fRoma	fEurope	fMiddle East	fSouth Asia
1:24193852-24197184	3332	rs7552167	none	IFNLR1;LINC02800	none	Psoriasis vulgaris;Psoriasis;Chronic inflammatory diseases (ankylosing spondylitis - Crohn's disease - psoriasis - primary sclerosing cholangitis - ulcerative colitis) (pleiotropy)	0,225	0,227	0,367	0,115
1:25405601-25409453	3852	rs10903129	RHCE	MACO1	none	Cholesterol - total;Total cholesterol levels;Erythrocyte sedimentation rate;LDL cholesterol levels;Low density lipoprotein cholesterol levels	0,688	0,364	0,533	0,525
1:26133625-26138449	4824	rs12088601	none	PDIK1L	none	Height	0,2	0,091	0,333	0,049
1:72290008-72298391	8383	rs61765646;rs34361149;rs2613499;rs2613498;rs61765651;rs2815749;rs2613500;rs2613503	none	RPL31P12;NEGR1	none	Insomnia symptoms (never/rarely vs. sometimes/usually);Gastroesophageal reflux disease;Body mass index;Childhood body mass index;Predicted visceral adipose tissue;Type 2 diabetes;Triglycerides;Heart rate response to recovery post exercise (10 sec);Childhood obesity;Adult body size	0,075	0,136	0,2	0,016
1:152582899-152615356	32457	rs6693105;rs11205044	LCE3B;LCE3C	LCE3B;LCE3A	LCE3B	Psoriasis or type 2 diabetes (trans-disease meta-analysis)(opposite effect);Chronic inflammatory diseases (ankylosing spondylitis - Crohn's disease - psoriasis - primary sclerosing cholangitis - ulcerative colitis) (pleiotropy)	0,562	0,523	0,6	0,533
1:169254682-169276856	22174	rs60486111	NME7	NME7	NME7	Myocardial infarction	0,212	0,273	0,367	0,148
1:174827380-174832728	5348	rs12564992	RABGAP1L	RABGAP1L	RABGAP1L	Body mass index	0,087	0,068	0,2	0,205
1:179360472-179364805	4333	rs12070589	none	TDRD5	none	Adolescent idiopathic scoliosis	0,05	0,045	0,033	0,033
1:185040669-185042601	1932	rs3806243	none	SWT1	none	Smoking behaviour (cigarettes smoked per day)	0,05	0	0,033	0,213
1:190868192-190868363	171	rs2053302	none	LINC01720	none	Response to statin therapy	0,75	0,864	0,967	0,852
1:209904708-209912632	7924	rs4844913;rs6661316	none	UTP25;SYT14	none	Body mass index;HDL cholesterol and triglyceride levels	0,512	0,614	0,633	0,459
1:229676721-229685143	8422	rs6700582	none	URB2;HMGB1P26	none	Height	0,438	0,409	0,467	0,385
2:34470755-34511503	40748	rs10495822	none	LINC01320	none	Amyotrophic lateral sclerosis (sporadic)	0,3	0,386	0,467	0,148
2:99487249-99488569	1320	rs3792142;rs6715321	REV1	AFF3;REV1	REV1	Self-reported math ability (MTAG);Highest math class taken (MTAG)	0,45	0,341	0,367	0,344
2:159103014-159104885	1871	rs448513	TANC1	TANC1	TANC1	Colorectal cancer or advanced adenoma	0,525	0,659	0,533	0,467
2:207486350-207494645	8295	rs13027546;rs13010104	none	CREB1;LRP1B	none	Refractive error;Cognitive ability - years of educational attainment or schizophrenia (pleiotropy)	0,225	0,205	0,267	0,074
3:15803568-15805240	1672	rs7612295	ANKRD28	BTD;ANKRD28	ANKRD28	Platelet count	0,1	0,045	0,033	0,18
3:37936926-37945436	8510	rs77388365;rs9816588	CTDSPL	CTDSPL	CTDSPL	Hemoglobin concentration;Hemoglobin;Lymphocyte counts;Hematocrit	0,138	0,068	0,2	0,107
3:41295190-41295738	548	rs9816029;rs9867802	ULK4	ULK4	ULK4	Body mass index;Predicted visceral adipose tissue	0,25	0,25	0,267	0,361
3:47449170-47451959	2789	rs8180040;rs13059519;rs17784882	SCAP	ELP6;PTPN23;KLHL18	none	Reaction time;QT interval;Colorectal cancer	0,6	0,409	0,333	0,377
3:136302170-136307360	5190	rs77273313;rs11279831;rs10935182;rs940174;rs10935186;rs6791142;rs7618871;rs1729951	PCCB	TAG1-DT;STAG1;MSL2;S	PCCB	Schizophrenia vs autism spectrum disorder (ordinary least squares (OLS));Neuroticism conditioned on self-rated math ability (multi-trait conditioning and joint analysis);Neuroticism;Schizophrenia;Feeling nervous;Anorexia nervosa - attention-deficit/hyperactivity disorder - autism spectrum disorder - bipolar disorder - major depression - obsessive-compulsive disorder - schizophrenia - or Tourette syndrome (pleiotropy);Neuroticism conditioned on Townsend deprivation index (multi-trait conditioning and joint analysis);Feeling worry;General risk tolerance (MTAG);Schizophrenia (MTAG);Neuroticism	0,487	0,341	0,433	0,451
3:191346881-191353900	7019	rs980320;rs980321	CCDC50	CCDC50	CCDC50	Gamma glutamyl transpeptidase;Gamma glutamyl transferase levels	0,2	0,159	0,2	0,18
3:197207702-197213386	4684	rs12495674	DLG1	DLG1	DLG1	Waist-hip ratio	0,287	0,25	0,367	0,221
4:10209636-10232949	23313	rs6839820	none	RAF1P1	none	Gout	0,7	0,568	0,667	0,664
4:107141930-107154888	12958	rs76067940	none	DKK2	none	Cortical surface area;Balding type 1;Male-pattern baldness	0,05	0	0,033	0,008
5:15719356-15720841	1485	rs61026653	FBXL7	FBXL7	FBXL7	Haemorrhoidal disease	0,125	0,114	0,1	0,213
5:24778170-24785918	7748	rs7735176	none	CDH10;BTG4P1	none	Polycystic ovary syndrome (reproductive subtype)	0	0,045	0	0,008
5:45004692-45014426	9734	rs62369897;rs62369904;rs6885307	none	HHC11B;NA;HCN1;MRPS	none	Height;Age at first birth	0,163	0,227	0,3	0,189
5:46270445-46275746	5301	rs4276421;rs10039283	none	none	none	PR interval;P wave duration	0,625	0,5	0,667	0,393
5:58027649-58037995	10346	rs62368409;rs1844437	none	LINC02101	none	High light scatter reticulocyte percentage of red cells;Hoarding	0,775	0,795	0,867	0,934
5:106989049-106990752	1703	rs4395620	LINC01950	LINC01950	LINC01950	Waist-hip ratio	0,175	0,364	0,267	0,426
5:127836638-127837588	950	rs245201	CCDC192	CCDC192	CCDC192	Brain imaging in schizophrenia (dorsolateral prefrontal cortex interaction)	0,575	0,659	0,533	0,656
5:178921367-178926196	4829	rs12186577	ZFP2	ZFP2	ZFP2	Refractive error;Spherical equivalent	0,175	0,114	0,1	0,008
6:19040878-19049239	8361	rs1322537;rs975303;rs72828517;rs72828532;rs72828533	none	CDK19;NA	none	Automobile speeding propensity;Sleep duration;Educational attainment (years of education);Self-reported math ability (MTAG);Intelligence (MTAG)	0,125	0,136	0,067	0,066
6:121476297-121476908	611	rs9490306	none	RNU4-35P;RNU4-76P	none	Migraine	0,25	0,182	0,167	0,27
6:131096843-131097847	1004	rs55898152	none	AKAP7;EPB41L2	none	Hip circumference adjusted for BMI	0,287	0,25	0,1	0,139
7:28174681-28175045	364	rs702814;rs849138;rs849142;rs849135	JAZF1	JAZF1	JAZF1	Systemic lupus erythematosus;Eye color;Type 2 diabetes;Systolic blood pressure;Sunburns;Monocyte count;Psoriasis or type 2 diabetes (trans-disease meta-analysis)	0,4	0,5	0,467	0,27
7:90181073-90183393	2320	rs39204	STEAP2-AS1	STEAP2-AS1	STEAP2-AS1	Type 2 diabetes	0,487	0,477	0,433	0,287
8:11388036-11389698	1662	rs2736308	FAM167A-AS1	FAM167A-AS1	FAM167A-AS1	Bisphosphonate-related osteonecrosis of the jaw in cancer (i.v. bisphosphonate)	0,45	0,409	0,367	0,221
8:24287757-24293571	5814	rs11992342;rs58578562	none	ADAM28	none	Cleft palate;Surgical necrotising enterocolitis in extremely premature birth	0,075	0,045	0,033	0,148
8:25553158-25559119	5961	rs76833264	none	CDCA2	none	Waist-to-hip ratio adjusted for BMI	0,037	0,023	0,033	0,074
8:86176023-86181514	5491	rs80037907	none	SLC7A13;ATP6V0D2	none	Educational attainment (MTAG)	0,1	0,068	0,067	0,057
8:93060098-93064968	4870	rs1027730	none	C8orf87;MIR8084	none	Attention deficit hyperactivity disorder	0,087	0,091	0,333	0,164
9:22991732-22992242	510	rs10965545	none	none	none	Tissue-type plasminogen activator levels	0,087	0	0	0,041
9:93708072-93709182	1110	rs4743930;rs12553508;rs10992867;rs12683791	none	MIR4291;PHF2	none	Facial morphology (segment 3);Predicted visceral adipose tissue;Heel bone mineral density	0,412	0,295	0,4	0,607
9:110262266-110267689	5423	rs4135188	none	TXN	none	Mean corpuscular hemoglobin	0,212	0,205	0,1	0,254
10:4666325-4668350	2025	rs2165953	MANCR	LINC00705;MANCR	MANCR	Obesity-related traits	0,675	0,864	0,767	0,754
10:30957818-30962846	5028	rs67326273	ZNF438	ZNF438;NA	ZNF438	Pulse pressure	0,15	0,227	0,333	0,074
11:27286339-27286655	316	rs7484147;rs7123749	none	CCDC34;BBOX1-AS1	none	Heel bone mineral density	0,65	0,659	0,433	0,697
11:27635716-27636048	332	rs4416056	BDNF-AS	BDNF-AS;LINC00678	BDNF-AS	Weight	0,675	0,705	0,467	0,443
11:28985626-28991394	5768	rs7126178;rs7123034	none	LINC02742	none	Educational attainment (MTAG);Highest math class taken (MTAG)	0,662	0,568	0,667	0,516
11:48658642-48658853	211	rs16905753	none	OR4X1	none	Hypertension	0,05	0,136	0,1	0,082
11:65874620-65876148	1528	rs2234458;rs2303385	none	EFEMP2	none	Body mass index;Adult body size;Retinal detachment or retinal break	0,75	0,682	0,567	0,836
12:10254813-10255122	309	rs4764369	KLRD1	KLRD1	KLRD1	Umami taste perception in obesity with metabolic syndrome	0,037	0,045	0,033	0,066
12:55333425-55334399	974	rs60683621	none	OR6C70	none	Odorant perception	0,15	0,318	0,167	0,213
12:99400170-99408992	8822	rs10128864	ANKS1B	ANKS1B	ANKS1B	Smoking initiation (ever regular vs never regular) (MTAG)	0,175	0,159	0,133	0,016
13:37497683-37511510	13827	rs962462	none	LINC00547;LINC01048	none	Blood protein levels	0,438	0,409	0,167	0,246
13:48959166-48962488	3322	rs183078	none	FNDC3A	none	Apolipoprotein A1 levels	0,237	0,386	0,467	0,336
13:55367041-55367333	292	rs1370063	none	none	none	Fish- and plant-related diet	0,625	0,75	0,6	0,475
13:64768301-64770291	1990	rs66481224	none	C13orf42;NA	none	Self-reported math ability (MTAG)	0,138	0,318	0,233	0,164
13:66600648-66603289	2641	rs76876592	PCDH9	PCDH9	PCDH9	Educational attainment (years of education)	0,163	0,159	0,3	0,098
14:73079061-73080255	1194	rs61986932	RBM25	PAPLN	none	Red cell distribution width	0,013	0,045	0	0,008
15:49253652-49255703	3051	rs11632038	GALK2	GALK2	GALK2	Lung adenocarcinoma	0,188	0,273	0,3	0,115
15:72093331-72095851	2520	rs12050794;rs2929508	MYO9A	MYO9A	MYO9A	Urate levels;Metabolite levels (HVA/MHPG ratio)	0,787	0,636	0,8	0,574
15:77038205-77040383	2178	rs11737	none	TSPAN3	none	Lung function (FEV1/FVC)	0,637	0,818	0,667	0,443
16:11589873-11591473	1600	rs7198919	none	LITAF	none	QRS complex (Sokolow-Lyon)	0,325	0,432	0,233	0,279
16:58613019-58616394	3375	rs246192;rs4784053	CNOT1	NDRG4;CNOT1	CNOT1	Liver enzyme levels (alanine transaminase);Sex hormone-binding globulin levels adjusted for BMI	0,362	0,341	0,433	0,131
16:76134750-76134855	105	rs2288061;rs2288059	none	RPL18P13	none	Delta-5 desaturase activity	0,188	0,273	0,1	0,205
16:87794758-87797857	3099	rs73249112	none	KLHDC4;SLC7A5	none	Facial morphology (factor 4 - facial height related to vertical position of gnathion)	0,138	0,159	0,1	0,164
17:59792637-59793865	1228	rs21250879;rs1292061	VMP1	VMP1	VMP1	Multiple sclerosis;Cortical thickness;White blood cell count (basophil)	0,362	0,409	0,4	0,426
18:26991656-26992377	721	rs1436904;rs12458099	CHST9	AQP4-AS1;CHST9	CHST9	Breast cancer	0,362	0,318	0,267	0,287
18:40426412-40427036	624	rs4799541;rs73429770	none	none	none	Subcortical volume (MOSTest);Brain morphology (MOSTest)	0,113	0,159	0,167	0,467
19:12584021-12588333	4312	rs7247513	ZNF490	ZNF490	ZNF490	Bipolar disorder	0,637	0,795	0,6	0,328
19:37850539-37854887	4348	rs34908476	WDR87BP	ZFP30;ZNF781	none	Lung function (FEV1/FVC)	0,062	0,136	0,067	0,008
20:1609399-1613359	3960	rs11696739	SIRPB1	SIRPG;SIRPB1	SIRPB1	Lymphocyte percentage of white cells;Mitochondrial DNA levels;Mean platelet volume	0,812	0,568	0,733	0,721
20:15321070-15323206	2136	rs6135385	MACROD2	MACROD2	MACROD2	Self-reported math ability (MTAG)	0,263	0,091	0,033	0,164
21:15216002-15219153	3151	rs2064040	none	none	none	Motor composite score	0,475	0,455	0,267	0,336
22:23683430-23685351	1921	rs17629956	GUSBP11	GUSBP11	GUSBP11	Adolescent idiopathic scoliosis	0,075	0,136	0,367	0,008

Supplementary Table 4. Table summarizing deletions in LD with GWAS SNPs. Coordinates, size SNP ID, and genes intersected by the deletions and by SNPs are reported, as well as common intersected genes, GWAS trait