UAB Universitat Autònoma de Barcelona

Establishing linkage disequilibrium between SNPs and polymorphic inversions

Bachelor's degree final project

(Genetics degree)

Author Éric García Hoyos

Tutor Prof. Dr. Francisco José Rodríguez-Trelles Astruga

31. 05. 2025

Acknowledgments

I want to thank my family for their continuous support throughout this project. I also want to express my gratitude to my friends, especially Amparo, for hearing my doubts and looking for ways to solve them, and Víctor, for revising the writing of this document and improving its clarity. I am also particularly grateful to Francisco José Rodríguez for his guidance as my project supervisor, as well as to Marta Puig and Ricardo Emanuel Moreira for introducing me to the field of genomic analyses and for their invaluable support.

Keywords

Genomics, inversion, structural variant, segmental duplication, inverted repeat, mobile element insertion, linkage disequilibrium, tagSNP.

.....

Éric García Hoyos
Universitat Autònoma de Barcelona (UAB), 31. 05. 2025

Index

1 Introduction

Early on the 21st century, the Human Genome Project (HGP) international consortium published the first draft of the human genome, with the objective to sequence the whole genome of our species and identify all genes contained in it [1]. Since this landmark achievement, subsequent studies have been centred on generating full resolved chromosomal maps. The most updated versions of a human reference genome have been published by the Genome Reference Consortium (GRCh38.p14/hg38) [2] and the Telomere-to-Telomere (T2T-CHM13) Consortium [3]. The public availability of this data has allowed the study of genomic variation across human populations, often understood as the detection of single nucleotide variants (SNVs). Nevertheless, there exist other rearrangements in our genome named as structural variants (SVs) and defined as regions of DNA larger than 50bp showing a change in copy number or genomic location [4, 5].

The rapid increase and the publication of fully resolved human genomes has been possible due to the use of third generation sequencing technologies (ONT and PacBio). Also, given that these technologies generate larger reads, a bigger proportion of the genome is captured in each sequencing round, so SVs can be more easily resolved and identified [4]. Therefore, publications containing information on genotypes, genomic coordinates and other features for SVs are being released by specific projects, such as the Human Genome Structural Variation Consortium (HGSVC) [6]. Datasets provided by this resource have been considered in this study, since this research attains particularly to the study of inversions, a specific type of SV in which the orientation of a DNA segment is contrary to that in the reference genome. But, even considering the great advances made in the field of genomics, the study of polymorphic inversions (those with a minimum allele frequency above 1% in human populations) is still very challenging due to the presence, in most of them, of segmental duplications (SDs) [7]. These genetic elements correspond to blocks of homologous DNA greater than 1Kb and with >90% sequence identity [8], found in both extremes of inversions in opposite orientations and containing the breakpoints (BPs) of these SVs. Inversions flanked by these types of repetitive sequences are generated by non-allelic homologous recombination (NAHR) events (see Figure 1), as it has been widely proposed in literature [8–11]. Given the homology of SDs, inversions generated by NAHR are highly recurrent in our genome, so standard and inverted status can be found across human populations independently of the ancestry and the inversion event that generated the SV [7, 11].

However, there also exist inversions that are not flanked by SDs. These SVs must have been originated by alternative mechanisms including double or single-stranded DNA break

processes, such as non-homologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ) [9], as well as replication based mechanisms, such as fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) [9, 11]. Given that these rearrangements are not as frequent as NAHR events, inversions originated by these mechanisms are not so recurrent and the different alleles for the inversion status are less diversified.

Since inversions have been associated with repression of recombination and disruption of regulatory and coding regions [7], I became particularly interested in the study of these SVs. By knowing the different mechanisms that could originate inverted fragments in our genome, I wanted to validate recurrence among these SVs and explore its relationship with haplotype disruption.

This was done by evaluating the non-random association of alleles at different loci in a given population. In our genome, certain combinations of genetic variants occur together more frequently than what it would be expected if the loci were assorting independently. So, co-segregation of alleles in the same haplotype can be studied with linkage disequilibrium (LD) measures. Various indices have been proposed in literature [12–14] and the discussion on which is the best measure of association lies on if these calculations are independent of allele frequencies or if data population structure is considered, so that the different values obtained can be compared [14].

Measuring LD between adjacent variants can serve to find perfect tag variants, those exhibiting perfect association with alleles of SVs, so that a particular allele of an inversion consistently co-segregates with the same allele of nearby variants. This approach has been used as a method to validate inversions' genotypes (given the existing difficulty due to the presence of SDs and other repetitive sequences) [15] and discover causal SVs which could drive correlated significant SNPs found in genome-wide association studies (GWAS) [13].

1.1 Investigation framework

In this research, a collection of inversions has been tested. For each of these, LD has been calculated by using the squared correlation coefficient (r^2) among alleles of inversions and surrounding Single Nucleotide Polymorphisms (SNPs) (see Equation 1). The selection of the r^2 statistic to measure LD lies in its correction for allele frequencies [12] and its usage in previous studies [9, 15].

$$r^2 = \left(\frac{D}{\sqrt{p_1 \cdot p_2 \cdot q_1 \cdot q_2}} \right)^2 \tag{1}$$

Where:

- r^2 is the squared correlation coefficient.
- p_1, p_2 are the allele frequencies for standard and inverted orientations, respectively.
- q_1,q_2 are the allele frequencies for SNPs.
- D is the LD coefficient: $(p_1 \cdot p_2) \cdot (q_1 \cdot q_2) - (p_1 \cdot q_2) \cdot (p_2 \cdot q_1)$.

This was done to find tagSNPs (with alleles perfectly associated to the inversion status, r^2=1) and therefore probe that:

- Inversions flanked by SDs, as they are originated by NAHR, are underrepresented among inversions with at least 1 tagSNP.
- Inversions generated by alternative mechanisms are overrepresented, since haplotypes have not been disrupted by recombinational events.

Also, this investigation has been guided by the accomplishment of a series of objectives, that stand for:

1. Develop a bioinformatic pipeline to provide a single file with merged data containing information for inversions, SNPs and the corresponding LD between variants.
2. Evaluate LD depending on the genomic context of inversions and find inversions perfectly tagged by surrounding SNPs.
3. Determine whether the presence of specific genetic elements, at the ends of inversions, influence the association between alleles of genomic variants.

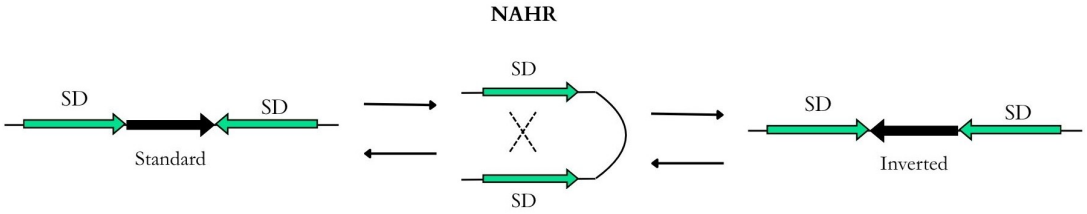

Figure 1: general architecture of an inversion flanked by SDs (marked in green). The inversion status can change depending on NAHR events, from standard to inverted and viceversa.

2 Methodology and Data Filtering

2.1 Workflow

A schematic representation of the methodology followed in this research is shown in Figure 2. The process is divided into different phases according to the organization in time. Each of these phases is explained in detail in the following sections.

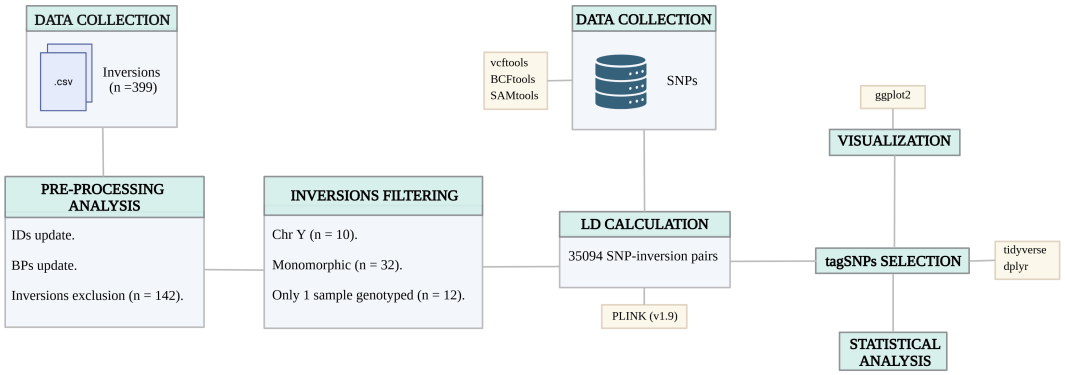

Figure 2: workflow for the different steps in which this research has been divided. Results coming from filtering steps are indicated. Green squares denote the different phases and orange squares mark different programs or packages used. Created using BioRender [16].

2.2 Data collection

2.2.1 Inversions dataset

Sequencing and genotyping information came from 3 different studies [7, 10, 17], although all data were consolidated in a single dataset [7]. Detailed information about the samples can be found in Supplementary Table A. In total, 399 inversions were analysed. Information for these SVs included their identification code (ID) and genomic coordinates for BPs (BP1 and BP2) in the GRCh38.p14/hg38 reference genome. These SVs ranged from 100 bp to 2Mb, although the average didn't exceed 50Kb.

A total of 44 samples were initially considered in this study. The selected data tried to have enough representation of different global superpopulations, shown in Figure 3. Absolute counts for samples were 7 EUR (European Ancestry), 13 AFR (African Ancestry), 6 SAS (South Asian Ancestry), 9 EAS (East Asian Ancestry) and 8 AMR (American Ancestry).

It is important to highlight that the sample NA24385 was not considered in this study (although it was included in the inversions' original dataset). That is because there was no available information in the 1000 GP datasets [6].

2.2.2 SNPs datasets

Information regarding genotypes, alleles and genome coordinates for SNPs in all 43 samples was extracted from the datasets provided by the online resource of the 1000 GP [18]. These variants were contained in all human chromosomes (except chromosome Y).

Figure 3: locations marked depending on the ancestry of the samples. Considered superpopulations were EAS (China, Japan and Vietnam in green), SAS (India, Pakistan and Bangladesh in blue), EUR (UK, Spain, Italy, Finland and Utah residents with Northern and Western European ancestry in red), AFR (Nigeria, Sierra Leone, Kenya, Gambia, African ancestry in South-West USA and Barbados in orange) and AMR (Colombia, Mexico, Peru and Puerto Rico in black).

2.3 Inversions filtering

2.3.1 Pre-processing analysis

After data collection, the next step consisted in the revision of information on genomic context and coordinates.

First, IDs for inversions contained in the original dataset were updated, so that these could be compared with already published studies. This was done by a search in the InvFest database [19], looking for the exact genomic coordinates of an inversion and consulting the ID specified for that SV. Some inversions remained with the old nomenclature, because no clear information was found. Then, after reviewing the genomic context for every inversion, a classification in different categories was made depending on flanking information

at the extremes of the SV. Two main categories were defined to probe the above-mentioned hypothesis (see section 1.1): inversions flanked by SDs (SD-flanked) and inversions where no SDs were identified (non-SD-flanked). Nevertheless, two additional categories were included in the analysis depending on if the inversion was flanked by inverted repeats (IR-flanked), similar to SDs but with lower sequence identity and less extended, or mobile elements insertion (MEI-flanked), including short and long interspersed elements (SINE and LINE, respectively), SINE-VNTR-Alu (SVA) elements, transposable elements with no distinction (TEs) and long tandem repats (LTRs).

Second, BPs for the inversions were initially set in a single point at each extreme of the SV, as shown in Figure 4. Nevertheless, given that inversions are not formed by the break in one single and invariant point, this information was updated to a wider range of nucleotides. The criteria to establish these depended on the inversion considered. The presence of repetitive sequences served to stablish these BPs. If inversions lacked these fragments, a range of 1000 bp was considered upstream and downstream the original BPs, considering all sites for excision and being representative enough.

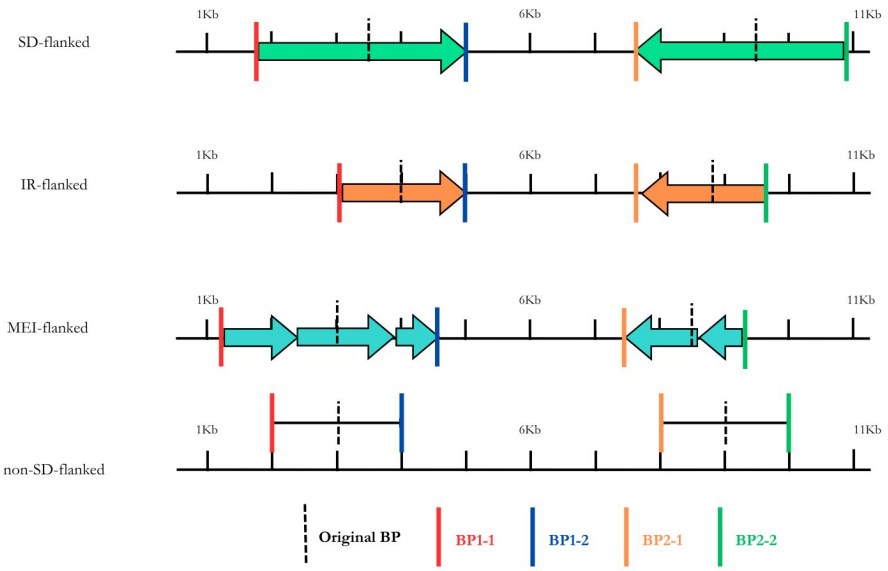

Figure 4: schematic view of the different inversions considered in this study. For each of them, the original BP is marked in a slashed line. Lines red (BP1-1) and blue (BP1-2) denote the extremes of the range considered to update the first BP, while lines orange (BP2-1) and green (BP2-2) mark the extremes of the range considered for the second BP. Arrows indicate the presence of repetitive elements: SDs in green, IRs in white and MEI in blue.

Finally, the collection of inversions was filtered by the criteria specified in Table 1, because of the possibility to generate biased or erroneous results when specific rearrangements (artifacts, duplicated, complex, errors, inverted dups, palindromic SDs and gaps) or chromosomal

locations (telomeric and centromeric) appeared. In total, 142 inversions were eliminated in this step, only 257 remained.

2.3.2 Genotypes information

Inversions were not only filtered by their information provided on genomic context, but also on the results coming from genotypes. It was excluded from the study a total of 32 monomorphic inversions (only 1 allele present in all samples for the inversion considered, either standard or inverted), 12 in which only 1 sample was genotyped (as it is necessary to have more then 1 sample to stablish LD) and 10 located in the chromosome Y (no available information in the 1000 GP datasets), among inversions where low confidence genotypes were not considered (Table 1).

To sum up, from the 257 remaining inversions after the pre-processing analysis, only 203 polymorphic inversions remained (54 were eliminated). The number of inversions for each flanking category among these 203 polymorphic inversions is indicated in Table 3, as well as the proportion of each.

Table 1: Exclusion of inversions that could generate errors or biased results. Thet total number for each category is specified in the second column (N). The description of every category is specified in the third column (Exclusion criteria).

Item	**N**	**Exclusion Criteria**
Total before filtering	399	
Artifacts	6	Overlaps another inversion.
Centromeric	17	Centromeric, with high content of repetitive DNA.
Complex	61	Overlaps other SVs.
Errors	2	Reference genome errors.
Inverted dups	44	Duplicated and inverted regions.
Telomeric	3	Telomeric, with high content of repetitive DNA.
Palindromic SDs	2	Overlap of SDs.
Gaps	1	Located in reference genome gaps.
Duplicated	6	Duplicate annotation.
Monomorphic	32	Only 1 allele.
Only 1 sample	12	Information for the genotype of only 1 sample.
Chr Y inversions	10	Inversions located in the chromosome Y.
Total after filtering	203	

2.4 LD calculation

After the filtering processes for inversions, the squared correlation coefficient (r^2) was used to stablish LD between every SNP-inversion pair by using PLINK (v1.9) [20]. This program required a window of bases to select specific variants in the inputs passed to it, so 500Kb downstream BP1-1 and upstream BP2-2 of the inversion were selected. This range considered a wide region to explore all SNPs contained in it, so that enough information was captured when searching for association between variants. Also, as another parameter of the software, only biallelic SNPs were considered.

This step was performed independently of samples' superpopulation and all data was treated as coming from one single group of individuals.

2.5 tagSNPs selection

In contraposition with inversions, SNPs were filtered once LD was already measured. Only those SNPs showing a value of r^2 equal to 1 were considered to be in perfect LD (tagSNPs), as it has already been mentioned in section 1.1.

2.6 Visualization

All the different results generated in this study were represented in informative graphs that came from R-based code (see used packages in section 2.8). A threshold of 0.35 for r^2 was considered when evaluating the distribution of LD measures (Figure 5) and also when evaluating the exact value of LD for every SNP-inversion pair (Supplementary Figure B).

2.7 Statistical analysis

Once all results were graphically represented, differences between groups were analysed. Given the format and the data generated, a two-sided Fisher's Exact Test was performed for every group of inversions to see if a specific category was significantly overrepresented or underrepresented. This test was done by comparing the total count of inversions in a specific category among the total of polymorphic inversions, and also comparing the total count of inversions for that category among inversions with at least 1 tagSNP.

Therefore, a 2x2 contingency table was constructed for every group of inversions (shown in Table 2). After the statistical test was performed, the corresponding p-value was adjusted using the Benjamin-Hochberg (BH) correction.

2.8 Packages, programs and code availability

Filtering and statistical analyses, as well as distribution evaluation, included different packages and were used in shell and R scripts, which have been uploaded to a Github repository [21], so they can be openly reviewed and accessed. The most relevant packages were tidyverse (v1.2.1) [22], dplyr (v1.1.4) [23] and ggplot2 (v3.5.2) [24].

LD was calculated using a series of shell scripts elaborated by an existing research, after requesting directly to the research group the availability of the code [15]. This was not uploaded to the Github repository because of privacity terms, but it uses the mentioned packages and others (BCFtools and SAMtools (v1.21) [25], vcftools (v0.1.17) [26]) to read information contained in Variant Calling Files (VCF), determine the desired range of bases (500Kb upstream and downstream the BPs, see section 2.5), process all data in binary files and calculate LD, as well as perform the appropriate transformations if the analysed inversion was located in the chromosome X (given that male individuals only have one allele).

Table 2: Contingency table for Fisher's Exact Test. x: inversions in a specific category with at least 1 tagSNP; y: polymorphic inversions in a specific category; N_{perf}: total with tagSNP; N_{poly}: total polymorphic.

	With tagSNP	**Polymorphic**
Category	x	y
Other inversions	$N_{\text{perf}} - x$	$N_{\text{poly}} - y$

3 Results

3.1 LD measures

After LD was measured for 203 inversions and surrounding SNPs, a series of measures were obtained. The majority of these ranged from 0.4 to 0.8 among all categories and a total of 35094 SNP-inversion pairs were above the cut-off stablished for the visualization (see section 2.6).

In Figure 5 it is represented the distribution of these values. Far from following a normal distribution, a great number of SNPs resulted to be in perfect LD. A shocking result is the great number of tagSNPs contained in inversions flanked by SDs, much greater than MEI-flanked and non-SD-flanked inversions.

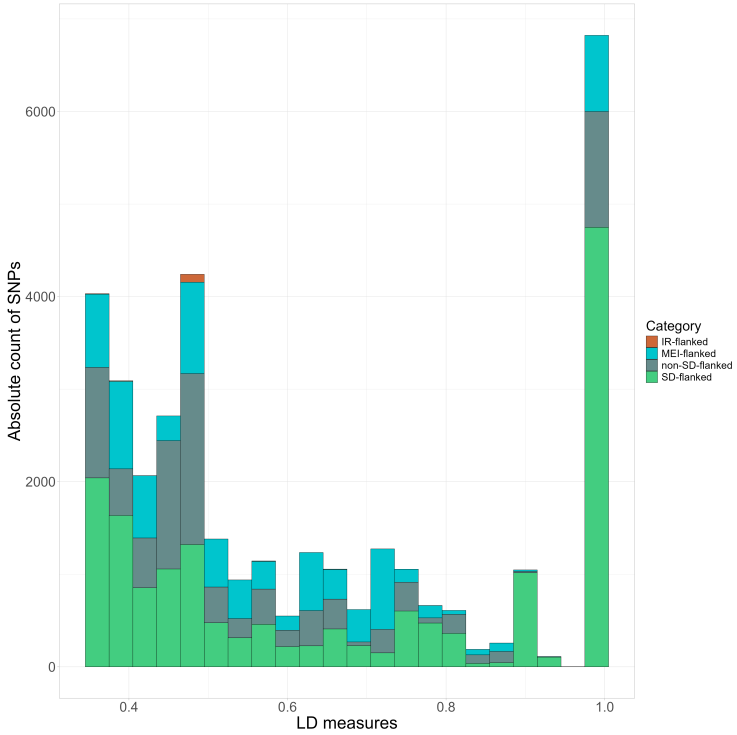

Figure 5: histogram showing the distribution of r^2 values. The y-axis refers to the absolute count of SNPs inside a determined range and divided depending on flanking information, while x-axis stands for r^2 values (from 0.35 to 1).

3.2 Inversions in perfect LD with SNPs

Only a specific number of inversions (86) were perfectly associated to alleles of SNPs. The corresponding number and proportion for each flanking category is indicated in Table 3. Some were surrounded by only 1 tagSNP, while others exhibited a greater number

of tagSNPs (see Figure 6). Unfortunately, the unique inversion categorized inside the IR-flanked category did not show perfect association with any surrounding variant.

Interestingly, a great number of SNPs (2873) in perfect LD corresponded to HsInv0573. This inversion, as it has been previously studied [27], is contained in a region where LD is very high and a lot of SNPs have been tagged, so these results are consistent with those expected. This result also explains the great number of SNPs obtained in perfect LD among SD-flanked inversions. The majority of these belonged to the region where HsInv0573 is located.

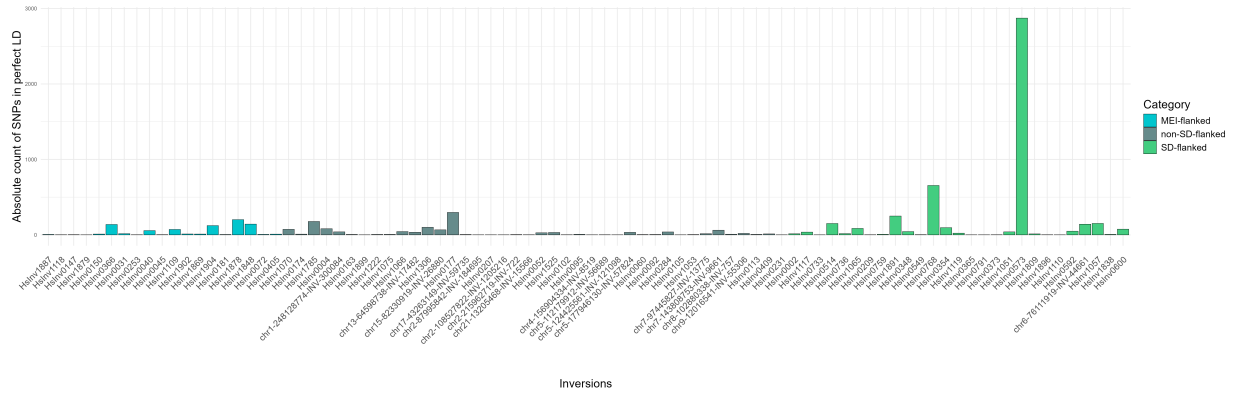

Figure 6: absolute count of tagSNPs (y-axis) divided by inversion (x-axis) and flanking information. Some of these inversions, as they had only 1 tagSNP, show shorter bars.

3.3 Proportions of inversions in perfect LD and statistical analysis

Representation of every flanking category among the total of polymorphic inversions and those with at least one tagSNP was also evaluated. Results for this analysis are indicated in Figure 7 and Table 3.

The statistical analysis performed (see section 2.7) generated the adjusted p-values shown in Table 3. It can be rapidly seen that IR-flanked inversions do not give any information on representation among groups of inversions. Also, non-SD-flanked inversions result to be significantly overrepresented (adjusted p-value of 0.043) among inversions with at least 1 tagSNP. On the contrary, SD-flanked inversions are significantly underrepresented among inversions in perfect LD (adjusted p-value of 0.0077). For inversions belonging to the the MEI-flanked category there were no significant differences (adjusted p-value of 0.326).

Table 3: proportions of inversion categories among polymorphic and tagSNP-containing inversions. P-values adjusted with BH correction. Significance: * (adj. p-value < 0.05), ** (adj. p-value < 0.01).

Metric	SD-flanked	Non-SD-flanked	MEI-flanked	IR-flanked	Total
Polymorphic count	107	62	33	1	203
Proportion in polymorphic	0.527	0.305	0.162	0.005	1
With tagSNP	28	39	19	0	86
Propoportion in tagSNP	0.326	0.453	0.221	0	1
Adj. p-value	0.0077	0.0429	0.3263	1	-
Significance	**	*	ns	ns	-

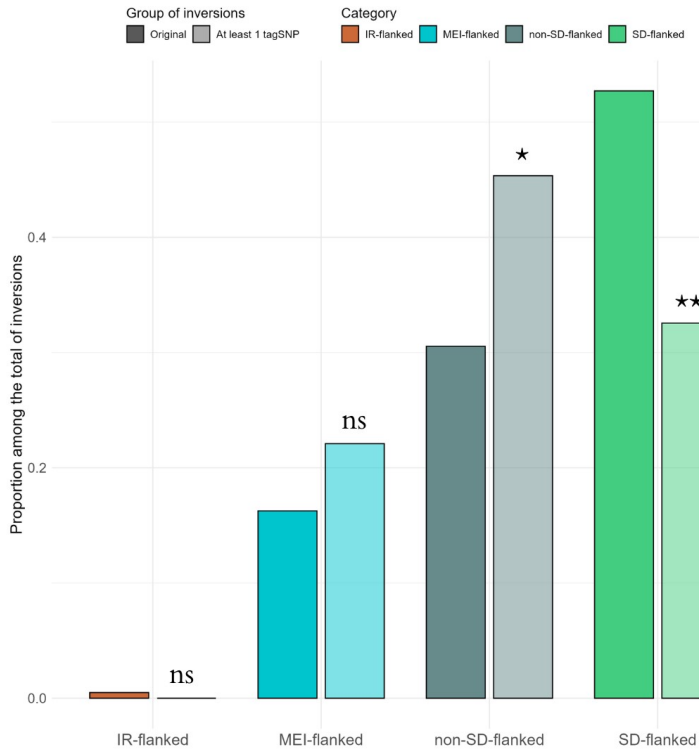

Figure 7: barplot representing proportion (y-axis) among polymorphic inversions (darker colours) and with at least 1 tag SNP (lighter colours). Four categories (x-axis) are divided depending on the genomic context. Asterisks located at the top of the bars indicate significance: * (adj. p-value < 0.05), ** (adj. p-value < 0.01).

3.4 Comparison with published data

LD measures were compared to those obtained in already published studies [9, 15]. Only for those inversions analysed both in this study and the above mentioned researches, the maximum LD value was selected for every SNP-inversion pair. Results for this validation are represented in Figure 8.

Among inversions where no IRs or SDs where identified, the great majority of inversions showed the presence of at least 1 tagSNP. Discrepancies were found only in HsInv0102 and HsInv0059, with overestimated and underestimated values of LD, respectively.

Inversions flanked by SDs and other repetitive sequences (IRs or MEI) showed a great variability in the data obtained, only being coincident for HsInv0040 and HsInv1051, these already associated with tagSNPs.

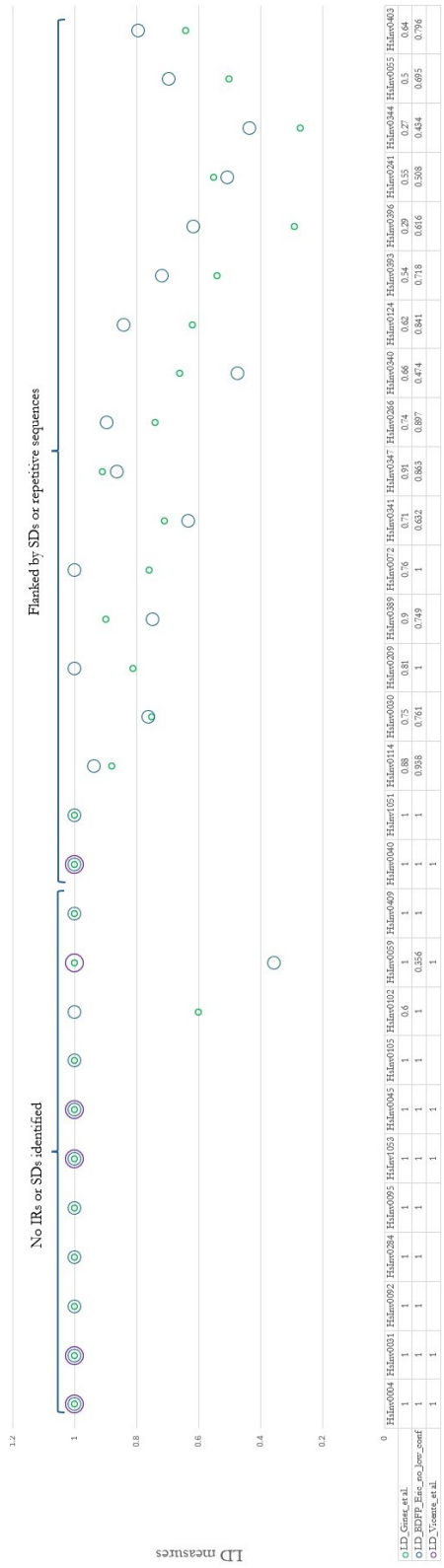

Figure 8: comparison of LD measures with already published data: LD_Giner et al. [15] and LD_Vicente et al. [9]. The different values obtained for every inversion are represented in a table format under the graphic. Also, inversions have been divided depending on the presence or absence of repetitive sequences.

4 Discussion and limitations

This study has been centred in the evaluation of LD between a set of polymorphic inversions and surrounding SNPs. After performing the appropriate filtering of the raw genotype and structural variant data, LD was measured to estimate the degree of association between these genomic variants. Results indicate that a significant lower proportion of inversions flanked in their extremes by SDs is in perfect LD with surrounding SNPs, while inversions in which no SDs have been identified are overrepresented among those with at least 1 tagSNP.

As it was noted in the section 1, the origin of polymorphic inversions in the human genome may result from diverse molecular mechanisms. Non-allelic homologous recombination (NAHR) is known to occur frequently between pairs of SDs due to their high sequence identity, often leading to recurrent inversions within the flanked region. For this, multiple and independent inversion events may occur across individuals, contributing to the haplotype disruption, so that even if a certain SNP is located very near the extremes of an inversion its alleles may not predict the orientation of the inverted fragment.

On the other hand, inversions generated by double or single-stranded DNA break processes (NHEJ and MMEJ) or replication based mechanisms (FoSTeS and MMBIR) are typically less recurrent. This is because these mechanisms do not rely on extensive sequence homology and are more likely to represent unique mutational events. So, if an inversion appeared relatively recently in the human genome by any of these mechanisms and it increased in frequency by any evolutionary force (such as genetic drift or natural selection) it would be associated with a single haplotype. Consequently, SNPs near the inversion would show strong LD with the inversion status, a reality consistent with the increased number of non-SD-flanked inversions found to be in perfect LD with at least 1 tagSNP in this study.

It is also worth mentioning that, although the results probe the initial hypothesis, certain inversions present unexpected results, as it is the case for HsInv0573 and HsInv0040, where a set of tagSNPs were encountered although these were classified as SD-flanked inversions. Also, by comparing the different r^2 values with already published data, some discrepancies were encountered (see section 3.4), specifically among those inversions flanked by repetitive sequences.

Because of this, some biases should be considered, possibly referred to:

- Data population structure, since all individuals have been considered as a single populational group, even knowing that allele frequencies and LD patterns can differ significantly between populations due to ancestral divergence, admixture, and selective

pressures.

- Reduced number of samples, only 43 samples (86 haplotypes) have been considered in this study.
- Bias in the LD measure, because the statistic r^2 is a good measure of association but is not totally independent of allele frequencies, so other measures considering haplotype based analyses could be implemented.

These limitations highlight the importance of cautious interpretation. Exceptions and noise within the data suggest that additional factors, such as local recombination rates, selective constraints, and genomic architecture, may also play critical roles.

Limitations aside, this project can serve as an approximation for genotypes imputation and validation when trying to characterize polymorphic inversions present in our genome. If a specific SV is associated with a tagSNP, alleles for the latter could be considered to validate the presence of the inversion in a genotyped individual. Even more, significant values for SNPs in Genome Wide Association Studies (GWAS) could be explained because of their perfect LD with an inversion (that is not directly analysed). This is of high interest because SVs (specifically inversions) can have substantial effects on gene expression, chromatin architecture, and regulatory landscapes [28], and may underlie associations previously attributed solely to SNPs. Thus, understanding LD between SNPs and polymorphic inversions is not only of theoretical interest but also of practical utility in medical genomics and evolutionary biology.

5 Conclusions

The different genetic variants studied in this research showed that the genomic context in which polymorphic human inversions are included directly affects LD measures with surrounding SNPs. Thus, NAHR events are responsible for haplotype diversification in SD-flanked inversions, while inversions that lack these genetic structures at their BPs could be originated in one single inversion event that left the inversion status perfectly associated ($r^2 = 1$) with SNPs' alleles, contributing to advances in genotype imputation and explanations on functional studies.

Also, the development of this research has generated a series of scripts to work with high-throughput data in the genomics field (see section 2.8). These are adressed to evaluate LD measures, inversions filtering and statistical analyses. This resource is of public access and can be found in a Github repository [21].

As a future perspective, this work could be complemented with studies considering larger sample sizes, populational structure biases and alternative LD measures to better understand association between genomic variants.

Bibliography

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409.** Erratum in Nature **412**,565(2001); Nature **411**, 720 (2001)., 860–921 (2001).
2. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27,** 849–864 (2017).
3. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376,** 44–53 (2022).
4. De Coster, W. & Van Broeckhoven, C. Newest Methods for Detecting Structural Variations. *Trends in Biotechnology* **37,** 973–982 (2019).
5. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
6. Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research* **48,** D941–D947 (2020).
7. Porubsky, D. *et al.* Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185,** 1986–2005.e26 (2022).
8. Jeong, H. *et al.* Structural polymorphism and diversity of human segmental duplications. *Nature Genetics* **57,** 390–401 (2025).
9. Vicente-Salvador, D. *et al.* Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Human Molecular Genetics* **26,** 567–581 (2017).
10. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372,** eabf7117 (2021).
11. Höps, W. *et al.* Impact and characterization of serial structural variations across humans and great apes. *Nature Communications* **15,** 8007 (2024).
12. Lewontin, R. C. On measures of gametic disequilibrium. *Genetics* **120,** 849–852 (1988).
13. Liang, H., Sedillo, J. C., Schrodi, S. J. & Ikeda, A. Structural variants in linkage disequilibrium with GWAS-significant SNPs. *Heliyon* **10,** e32053 (2024).
14. Mangin, B. *et al.* Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* **108,** 285–291 (2012).
15. Giner-Delgado, C. *et al.* Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature Communications* **10,** 4222 (2019).
16. BioRender. *BioRender.com* https://biorender.com. (2024).
17. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications* **10,** 1784 (2019).
18. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185,** 3426–3440.e19 (2022).
19. Martínez-Fundichely, A. *et al.* InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Research* **42,** D1027–D1032. http://invfestdb.uab.cat/search_invdb2.php (2014).

20. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81,** 559–575 (2007).

21. Hoyos, E. G. *Bachelor Degree Final Project* (2025). https://github.com/Eric1630394/Bachelor_degree_final_project.

22. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4,** 1686. https://doi.org/10.21105/joss.01686 (2019).

23. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *dplyr: A Grammar of Data Manipulation* (2023). https://dplyr.tidyverse.org.

24. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* https://ggplot2.tidyverse.org (Springer-Verlag, New York, 2016).

25. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10,** giab008 (2021).

26. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

27. Campoy, E., Puig, M., Yakymenko, I., Lerga-Jaso, J. & Cáceres, M. Genomic architecture and functional effects of potential human inversion supergenes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **377.** Epub 2022 Jun 13, 20210209 (2022).

28. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nature Genetics* **49,** 692–699 (2017).

A Supplementary Table

Table S1: Information on Sample ID, Sex, Superpopulation code, and Study of origin for the 43 samples analysed. Sequencing and genotyping information has come from three studies, two of them part of the Human Genome Structural Variation Consortium, Phase 1 and Phase 2 (HGSVC1 and HGSVC2, respectively): Chaisson et al. (HGSVC1) [17], Ebert et al. (HGSVC2) [10], and Porubsky et al. [7].

Sample ID	Sex	Superpopulation code	Study of origin
HG00096	male	EUR	Ebert et al. (HGSVC2)
HG00171	female	EUR	Ebert et al. (HGSVC2)
HG00268	female	EUR	Porubsky et al.
HG00512	male	EAS	Chaisson et al. (HGSVC1)
HG00513	female	EAS	Chaisson et al. (HGSVC1)
HG00514	female	EAS	Chaisson et al. (HGSVC1)
HG00731	male	AMR	Chaisson et al. (HGSVC1)
HG00732	female	AMR	Chaisson et al. (HGSVC1)
HG00733	female	AMR	Chaisson et al. (HGSVC1)
HG00864	female	EAS	Ebert et al. (HGSVC2)
HG01114	female	AMR	Ebert et al. (HGSVC2)
HG01352	female	AMR	Porubsky et al.
HG01505	male	EUR	Ebert et al. (HGSVC2)
HG01573	female	AMR	Porubsky et al.
HG01596	male	EAS	Ebert et al. (HGSVC2)
HG02011	male	AFR	Ebert et al. (HGSVC2)
HG02018	female	EAS	Porubsky et al.
HG02059	female	EAS	Porubsky et al.
HG02106	female	AMR	Porubsky et al.
HG02492	male	SAS	Ebert et al. (HGSVC2)
HG02587	female	AFR	Ebert et al. (HGSVC2)
HG02818	female	AFR	Ebert et al. (HGSVC2)
HG03009	male	SAS	Ebert et al. (HGSVC2)
HG03065	male	AFR	Ebert et al. (HGSVC2)
HG03125	female	AFR	Ebert et al. (HGSVC2)
HG03371	male	AFR	Ebert et al. (HGSVC2)
HG03486	female	AFR	Ebert et al. (HGSVC2)
HG03683	female	SAS	Ebert et al. (HGSVC2)
HG03732	male	SAS	Ebert et al. (HGSVC2)
HG04217	female	SAS	Porubsky et al.
NA12329	female	EUR	Ebert et al. (HGSVC2)
NA12878	female	EUR	Ebert et al. (HGSVC2)

Sample ID	Sex	Superpopulation code	Study of origin
NA18534	male	EAS	Ebert et al. (HGSVC2)
NA18939	female	EAS	Ebert et al. (HGSVC2)
NA19036	female	AFR	Porubsky et al.
NA19238	female	AFR	Chaisson et al. (HGSVC1)
NA19239	male	AFR	Chaisson et al. (HGSVC1)
NA19240	female	AFR	Chaisson et al. (HGSVC1)
NA19434	female	AFR	Porubsky et al.
NA19650	male	AMR	Ebert et al. (HGSVC2)
NA19983	female	AFR	Ebert et al. (HGSVC2)
NA20509	male	EUR	Ebert et al. (HGSVC2)
NA20847	female	SAS	Ebert et al. (HGSVC2)

B Supplementary Figure

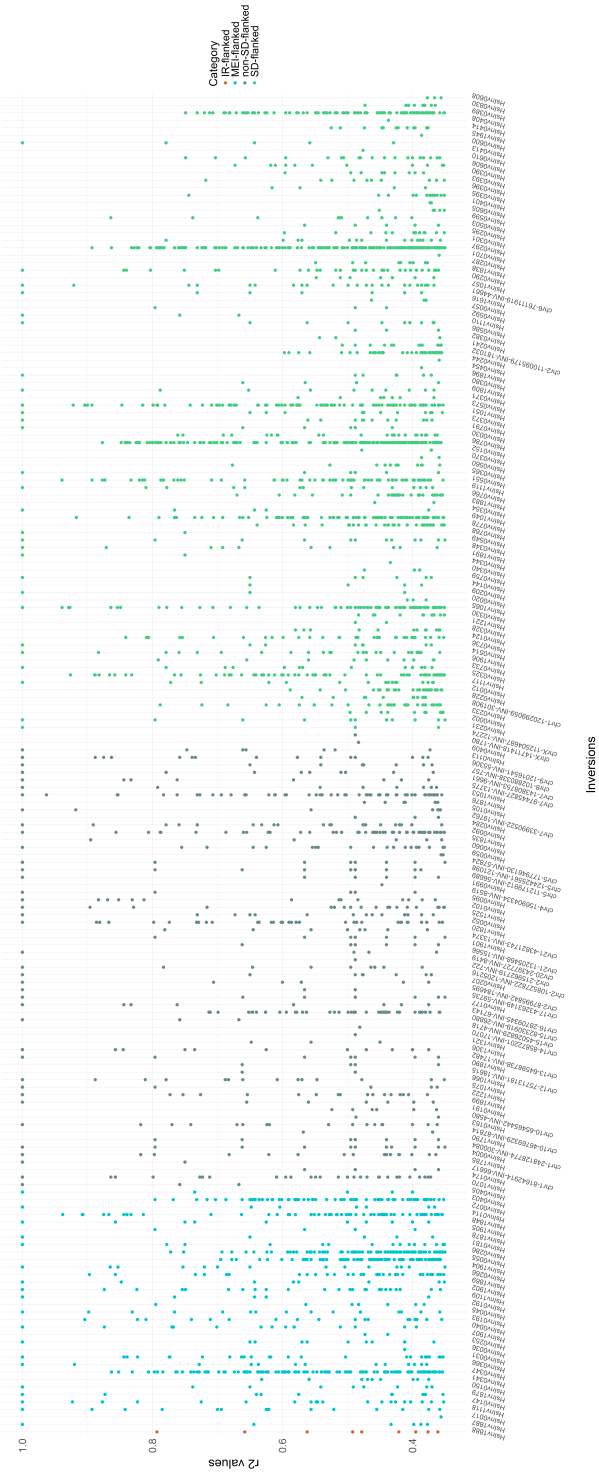

Figure S1: representation of r^2 values for every SNP-inversion pair. Each point corresponds to the r^2 value obtained for a specific SNP (y-axis) and the corresponding inversion (x-axis). Inversions have also been divided depenidng on flanking information.