HaploExplore, a software specifically designed for the detection of minor allele (MiA-) Haploblocks

Matilde Manetti[1]*, Samuel Hiet[1]*, Myriam Rahmouni[1], Jean-Louis Spadoni[1], Alice Dobiecki[1], Marco Lamanda[1], Maxime Tison[1], Taoufik Labib[1], Cristina Giuliani[2], Sigrid Le Clerc[1], Jean-François Deleuze[3], Jean-François Zagury[1]

[1] Laboratoire Génomique, Bioinformatique, et Chimie Moléculaire, EA7528, Conservatoire National des Arts et Métiers, 2 rue Conté 75003 – Paris, France.

[2] Laboratory of Molecular Anthropology, Department of Biological, Geological and Environmental Sciences, University of Bologna, Via Selmi 3, Bologna, Italy.
[3] Laboratory for Genomics, Foundation Jean Dausset – CEPH, Paris, France.

*The 2 first authors share an equal contribution for this work.

Corresponding author: Jean-François Zagury, zagury@cnam.fr

Supplementary Materials

1. Method

1.1 Default settings

The software supports adjustable parameters, including: LD thresholds (default r^2 =0.1, D′=0.7), carrier percentage cut-off (default 0.8), MAF percentage cut (default 0.8), maximum SNP gap within a block (default 200 SNPs), maximum haploblock size (default 5M base pairs), region size for splitting datasets (default 10M base pairs), while the minimum MAF threshold to consider a SNP is 1% (default value).

Carrier Percentage plays a crucial role in detecting haploblocks of minor alleles even when traditional linkage disequilibrium measures, such as r^2, are low (e.g., 0.1) but D′ remains high (e.g., 0.7). Unlike LD-based approaches that rely on strong statistical correlations, CP directly assesses the co-occurrence of minor alleles within individuals, ensuring that SNP2 is included in a haploblock with SNP1 when they frequently appear together. By focusing on the actual presence of alleles in individuals rather than just correlation values, CP allows the detection of weaker genetic associations that might otherwise be missed due to a low r^2, this is particularly useful when SNPs are in high D' but low r^2. The default value for the carrier percentage threshold has been set at 80%. This choice is confirmed by both biological and statistical considerations. The haploblock should include SNPs whose minor alleles sufficiently cover the coreSNP minor allele to help explain the genetic association detected in statistical analyses. An 80% carrier percentage ensures that SNPs in the haploblock are strongly linked to the coreSNP. If a causal SNP is located within the haploblock of a coreSNP identified through statistical screening, we expect that the majority of individuals carrying the coreSNP minor allele also carry the minor allele of the causal SNP.

The default CP threshold (80%) was chosen based on empirical observations from the European haploblock containing SNP rs2395029, a known marker for HLA-B57:01 which is important for Elite Control in AIDS. In this setting, CP 80% corresponds to LD values of approximately $r^2 > 0.1$ and D′ > 0.7. Users may adjust this threshold depending on their aim, generally by relaxing the numbers they will get more SNPs and longer haploblocks, and by constraining the parameters they will get the opposite (Supplementary Table S1). If one wants to focus more on haploblock structure (identification of genetic domains) you will try to work more with D' and r^2, but CP will remain an important parameter since it allows to put the focus on MiA-haploblocks. If one wants to focus more on the biological activity of SNPs, one will relax more D' and r^2, and keep CP constrained.

Setting the haploblock MAF percentage cut ($\alpha_{cut\,\%}$) at 0.8 of the coreSNP's MAF ensures that only SNPs with a MAF of at least 80% of the coreSNP's MAF are considered for inclusion. This allows SNPs with slightly lower frequencies than the coreSNP to still be part of the haploblock and contribute to explaining the statistical impact associated with the coreSNP. For example, if the coreSNP has a MAF of 0.4, only SNPs with a MAF of at least 0.32 will be evaluated for inclusion. If SNPs are pre-filtered with a MAF threshold of 0.01, then both the coreSNP and the selected SNPs will have a minimum MAF of 0.01. This ensures that haploblocks are not constructed with extremely rare variants, which could otherwise produce unstable or non-representative blocks. If there is no MAF pre-filtering, the MAF percentage cut (e.g., 80%) will allow inclusion of rarer variants; for example, if the core SNP has MAF = 0.01, then selected SNPs may have MAF as low as 0.008. This flexibility allows users to explore haploblocks that include SNPs with lower MAF than the core SNP, but with a minor allele highly co-inherited with the minor allele of the core SNP.
From our preliminary tests, the maximum SNP gap value of 200 seems to be a reasonable value which allows us to generate more compact haploblocks.
For efficiency reasons, the genome is cut in successive parts to explore the haploblocks, and we observed that the optimal region size to cut the genome was twice the maximal haploblock size.
We have performed a sensitivity analysis to examine the effects of varying CP, r^2, and D′ thresholds on haploblock detection. The resulting haploblock size distributions and SNP count under different parameter settings are shown in Supplementary Table S1 and Supplementary Figure S1. These results demonstrate that users who recover either too many or too few signals (in terms of either number of haploblocks or number of SNPs in a haploblock) can readily adjust the parameters to modulate signal detection according to their research objectives. Although it is not possible to qualify the default parameters as strictly optimal, the analyses indicate that they constitute a very reasonable and practical baseline.

Table S1 Varying one parameter at a time (Defaults: CP = 0.8, r^2 = 0.1, D′ = 0.7, MAF cutoff = 0.8) summary of Number of haploblocks, average snps per block, and average size in bp. The haploblocks were computed from 500 individuals of the DESIR cohort in the full chromosome 22.

Parameter	Value	Number of haploblocks	Mean in SNPs	Mean in bp
CP	0.50	9680	26.1	75,025
	0.70	9680	26.0	74,652
	0.80	10298	25.3	70,692
	0.90	12936	20.7	53,761
r^2	0.1	10298	25.3	70,692
	0.3	14956	14.0	43,013
	0.5	21071	9.48	30,297
D′	0.5	10183	25.8	72,756
	0.7	10298	25.3	70,692
	0.9	14274	19.0	48,399

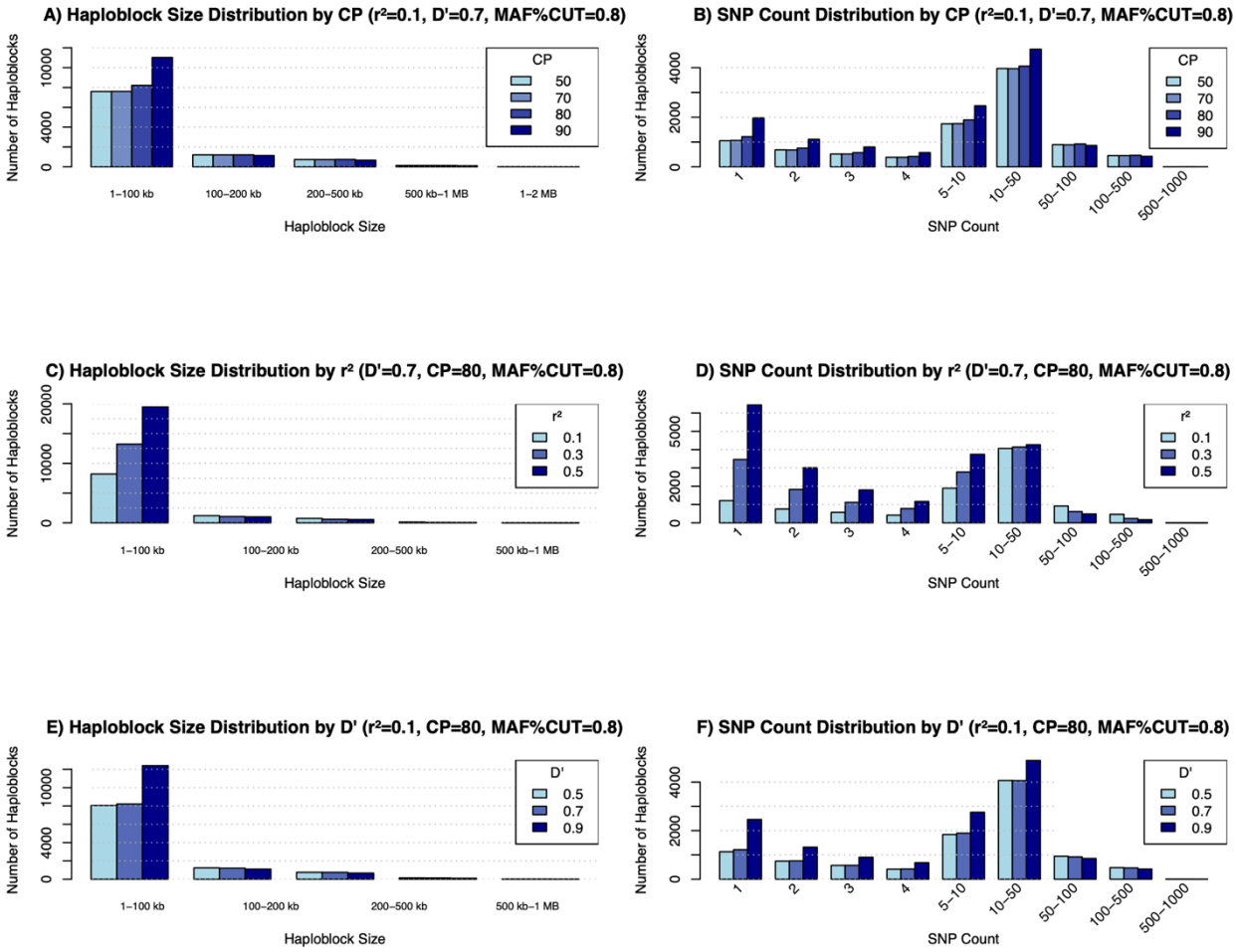

Figure S1. ***Effect of HaploExplore parameters on haploblock size and SNP count distributions.*** *Haploblocks were computed from 500 individuals of the DESIR cohort in the full chromosome 22. Barplots show the number of haploblocks categorized by (A) haploblocks size (in bp) vs. CP, (B) SNP count vs. CP, (C) haploblocks size vs. r², (D) SNP count vs. r², (E) haploblocks size vs. D', and (F) SNP count vs. D'. Each panel highlights the influence of varying a single parameter while fixing the others (default values: r² = 0.1, D' = 0.7, CP = 80, MAF cutoff = 0.8). Colors represent different parameter thresholds, and dotted grid lines facilitate comparison across haploblock categories.*

Panels (A) and (B) show that increasing CP slightly reduces the number of large haploblocks (>200 kb), while panels (C)–(F) indicate that stricter LD thresholds (r² or D′) decrease the proportion of large haploblocks and increase the proportion of smaller blocks (1–100 kb). The MAF cutoff primarily acts as a pre-filter and has minimal effect on haploblock size and SNP count distributions; therefore, plots for varying MAF thresholds are not shown.

Table S2: *Comparison of haploblock size (number of SNPs) generated using the parameters r^2, D', and CP, either individually or in combination, for the core SNP rs2395029 (minor allele frequency: 4%) located on chromosome 6, based on 500 individuals from the DESIR cohort. Each cell reports the proportion (%) of SNPs shared between two haploblock definitions, with the corresponding number of overlapping SNPs relative to the total number in the reference haploblock indicated in parentheses.*

	r^2 (0.1)	**D' (0.7)**	**CP (0.8)**	**Default (CP = 0.8; D′ = 0.7; r^2 = 0.1)**
r^2 (0.1)	100.00 (662/662)	91.24 (604/662)	65.56 (434/662)	65.56 (434/662)
D'(0.7)	4.46 (604/13541)	100.00 (13541/13541)	14.39 (1949/13541)	3.21 (434/13541)
CP (0.8)	22.02 (434/1971)	98.88 (1949/1971)	100.00 (1971/1971)	22.02 (434/1971)
Default (CP = 0.8; D′ = 0.7; r^2 = 0.1)	100.00 (434/434)	100.00 (434/434)	100.00 (434/434)	100.00 (434/434)

Table S3. Summary statistics of haploblocks (DESIR cohort, *n* = 500 individuals, in the full chromosome 22 with haploblocks of minimum size 10 SNPs). Haploblocks were computed under different parameter configurations.

Setting	**Number of haploblocks**	**Mean in SNPs**	**Mean in bp**
CP (0.8)	6,753	121.55	237,751
D′ (0.7)	1,488	313.19	386,989
r^2 (0.1)	4,537	58.27	136,664
Default (CP = 0.8; D′ = 0.7; r^2 = 0.1)	5,446	44.14	106,139

1.2 Output Data

The software proposes several types of graphical outputs such as histograms (SNP count, size in BP…) as shown by the figures provided in the main text. Additionally, it also proposes the visualization of haploblocks in the chromosome as shown in Figure S2 below.

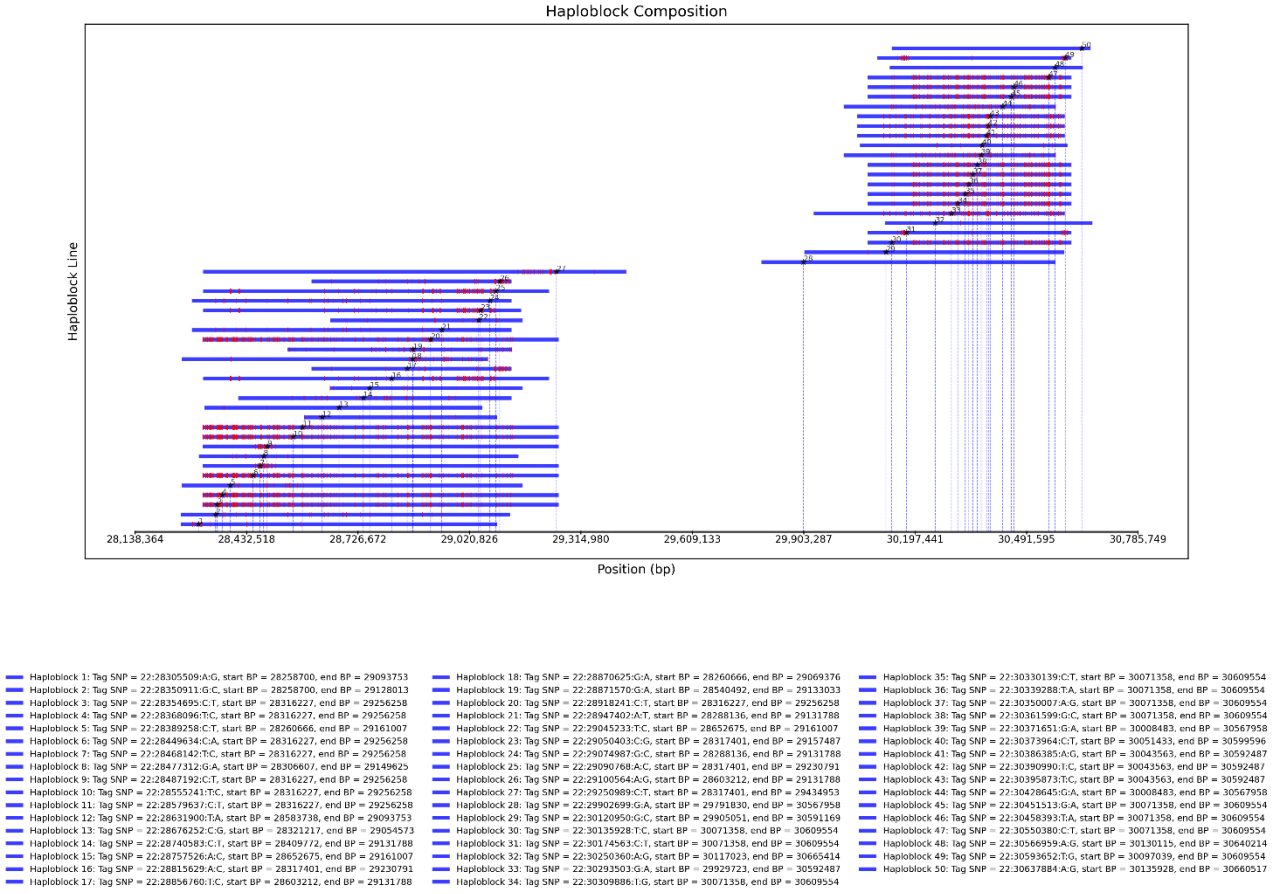

Figure S2 Graphical Representations of Haploblock Boundaries Using List SNPs Mode. *Haploblock boundaries are identified using the List SNPs mode. There were 1000 SNPs in the predefined SNP list in chromosome 22, and they serve as coreSNP to build new haploblocks and are marked by a star. SNPs in red indicate those with a Minor Allele Frequency within ±10% of the MAF of the corresponding coreSNP (to visualize some kind of haplotype). The remaining SNPs within the haploblock are depicted in blue. This representation highlights the relationship between coreSNPs, surrounding SNPs, and the overall haploblock structure. The x-axis represents the SNP positions in base pairs (bp), while the y-axis corresponds to haploblock lines. The software creates more pdf files, for each pdf it creates a graph with maximum 50 haploblocks, the haploblocks are plotted according to the order of the list when they are generated, this is the first plot that we obtain from the analysis.*

2. Results

2.1 Convergency

To assess the stability of haploblock detection across different sample sizes, we analyzed the number of haploblocks identified and the corresponding genomic coverage as the sample size increased for the Standard mode with default parameters. The results indicate a clear trend toward convergence, with diminishing changes beyond 100–250 individuals.

Table S4 Impact of sample size on haploblocks detection: number of haploblocks identified across different sample sizes of genotyped individuals of the DESIR cohort in the full chromosome 22, with relative change indicating the increase in haploblock count as sample size grows

Sample Size	Haploblocks	Relative Change
25	4,660	-
50	5,985	+22.1%
100	9,655	+38.0%
250	10,031	+3.7%
500	10,295	+2.5%

As seen in Table S4, the total number of haploblocks increases substantially as sample size grows, particularly between 25 and 100 individuals, where the number more than doubles (+107%). However, beyond 100 individuals, the increase becomes much more gradual, with only a 3.7% change from 100 to 250 samples and 2.5% from 250 to 500 samples. This suggests that after a certain threshold, the software stabilizes in its haploblock detection capabilities.

Table S5 Effect of sample size on genomic coverage, displays the total genomic coverage (in base pairs) for various sample size of genotyped individuals from the DESIR cohort in the full chromosome 22, with relative change indicating the variation in coverage as sample size increases

Sample Size	Coverage (bp)	Relative Change
25	34,096,342	-
50	34,131,043	+0.10%
100	33,890,252	-0.71%
250	32,834,246	-3.2%
500	31,802,966	-3.2%

In contrast, Table S5 shows that genomic coverage remains relatively stable at smaller sample sizes but starts to decrease as the sample size increases beyond 100 individuals. The reduction in coverage suggests that the detection process is refining haploblocks, leading to finer resolution but also potentially filtering out larger blocks that may still be biologically meaningful. While larger sample sizes allow for the detection of more haploblocks, many of these additional haploblocks tend to be smaller in size. From a biological perspective, excessively small haploblocks may result from noise, sequencing artifacts, or minor variations in recombination events, making them less relevant for functional analysis. Thus, it is crucial to find a balance between accuracy and resolution—ensuring robust haploblock detection while minimizing false positives and excessive fragmentation. Based on our findings, a sample size of 100–250 individuals appears to offer an optimal trade-off: At this range, haploblock detection is stable. Larger haploblocks remain well-defined. The risk of excessive, biologically uninformative small haploblocks is minimized. Beyond 250 individuals, the increase in detected haploblocks is minor, but the proportion of smaller haploblocks rises, potentially introducing noise rather than improving accuracy. Therefore, for most

studies, a sample size of around 100–250 individuals may be the most efficient choice for haploblock analysis.

Table S6 Average haploblock size according to the sample size of genotyped individuals from the DESIR cohort in the full chromosome 22: average number of SNPs per haploblock for different sample sizes, showing how haploblock size varies with increasing sample size.

Sample Size	Mean size Haploblocks (# SNPs)
25	33.54
50	35.48
100	27.16
250	25.49
500	25.20

2.2 Impact of population size

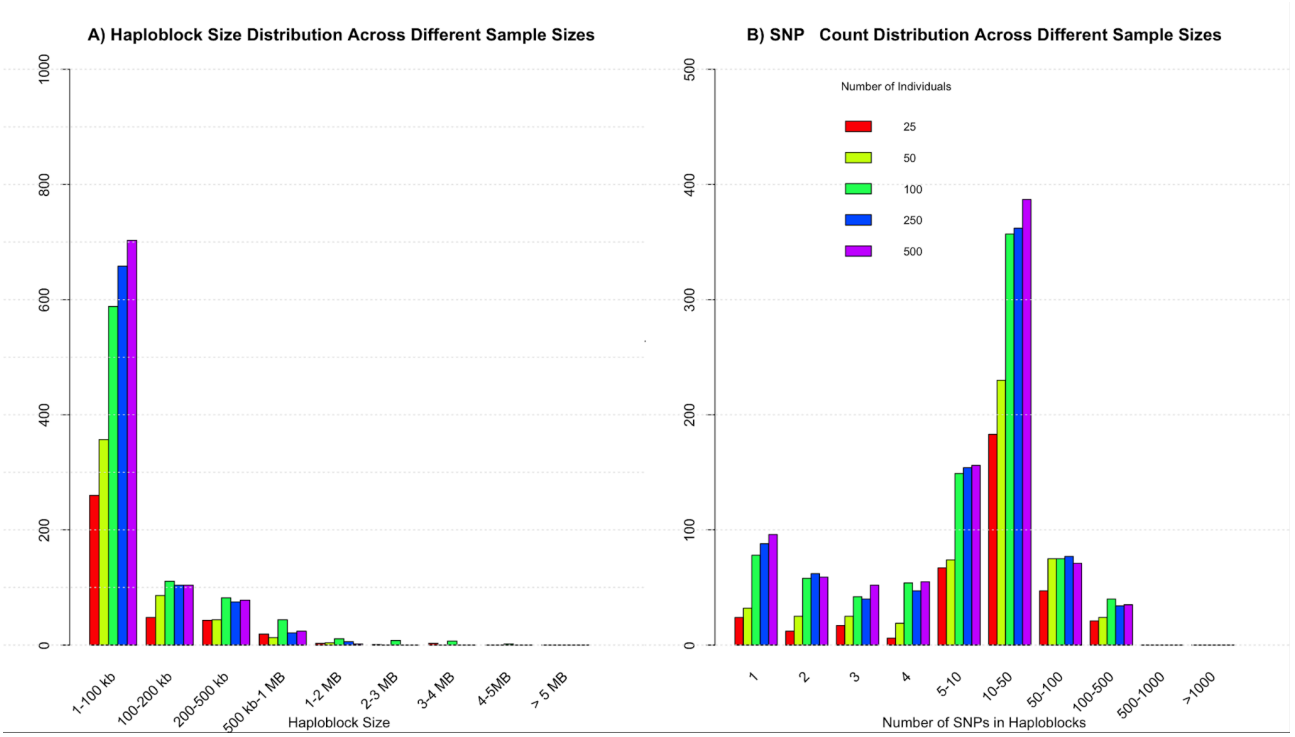

Figure S3 (A) *Distribution of haploblock sizes across various sample sizes from the DESIR cohort* in the full chromosome 22. *The x-axis displays haploblock size categories, ranging from 1–100 kb to over 5 MB, while the y-axis represents the number of haploblocks observed in each size category. Bar plots are used to show the distribution for each sample size (25, 50, 100, 250, 500 individuals), with distinct colors representing different sample sizes. The results highlight how larger sample sizes (250–500 individuals) lead to a more stable and reliable representation of haploblock size distributions.* **(B)** *Distribution of SNP counts per haploblock across various sample sizes in the analysis. The x-axis shows SNP count categories (ranging from 1 to >1000 SNPs per haploblock), while the y-axis represents the number of haploblocks observed. Bar plots are used to display the results for all sample sizes (25, 50, 100, 250, 500 individuals), with each sample size represented by a unique color. Both panels are based on data from chr 22 using the standard mode of HaploExplore. Parameters include: LD thresholds ($r^2 \geq 0.1$, $D' \geq 0.7$), MAF ≥ 0.8, %carrier threshold ≥ 80, region size of 10,000,000 base pairs with a region overlap of 5,000,000 base pairs, and a maximum empty gap of 200 SNPs.*

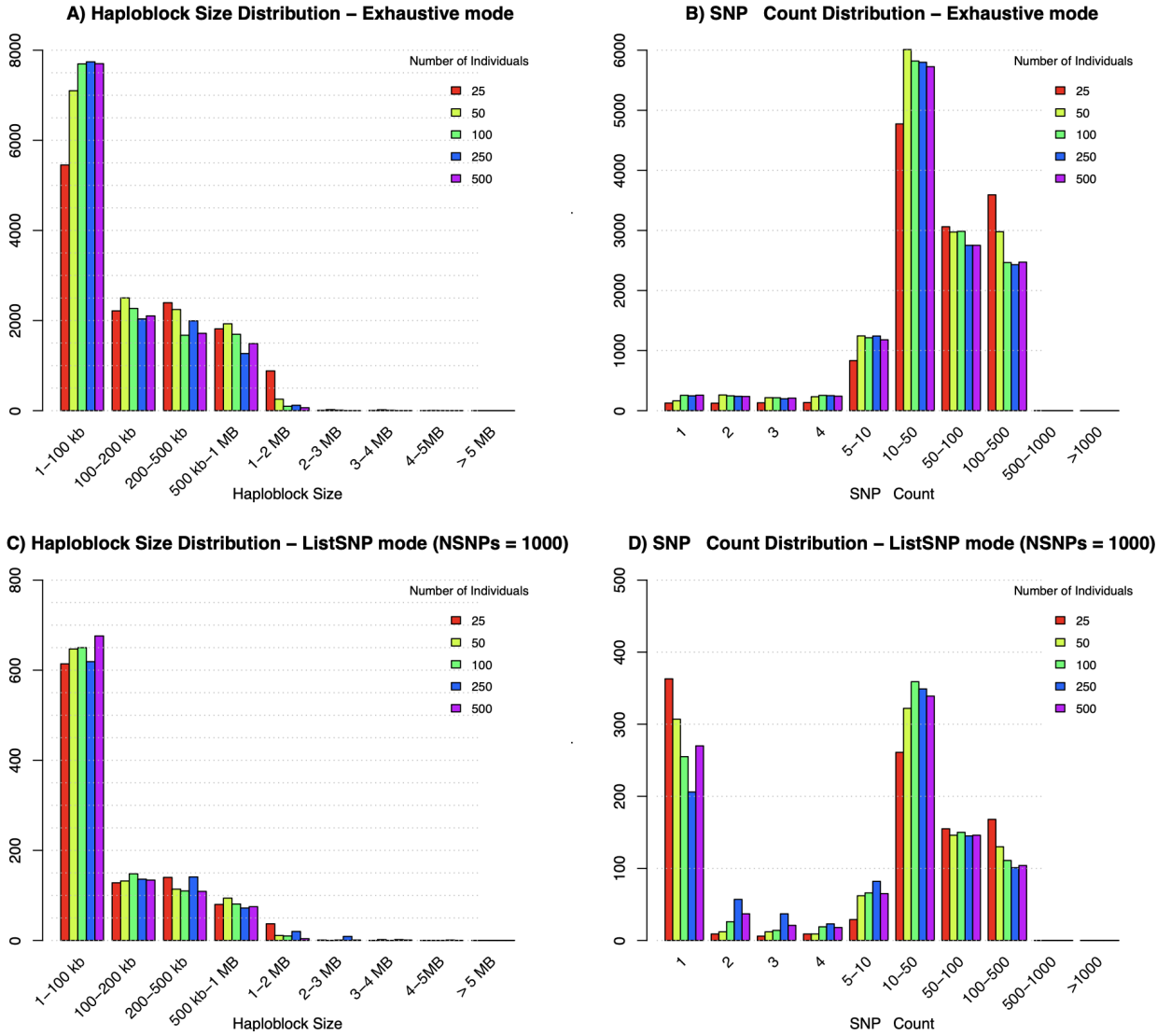

Figure S4 Haploblock and SNP Count Distributions for Different Modes. (A) *Haploblock size distribution using the Exhaustive mode.* **(B)** *SNP count distribution using the Exhaustive mode.* ***(C)*** *Haploblock size distribution using the List SNPs mode.* **(D)** *SNP count distribution using the List SNPs mode, with the number of SNPs in the predefined list set to 1,000. These bar plots illustrate how haploblock sizes and the number of SNPs per haploblock vary with increasing sample sizes (25, 50, 100, 250, 500 genotyped individuals of the DESIR cohort* in the full chromosome 22*). Each plot uses a smaller genomic region of 5 Mb of chromosome 22 for the analysis, with parameters: LD thresholds ($r^2 \geq 0.1$, $D' \geq 0.7$), MAF ≥ 0.8, %carrier threshold ≥ 80, region size of 5,000,000 base pairs, and a maximum empty gap of 200 SNPs. These visualizations highlight the differences between modes and how sample size impacts haploblock structure and SNP distribution.*

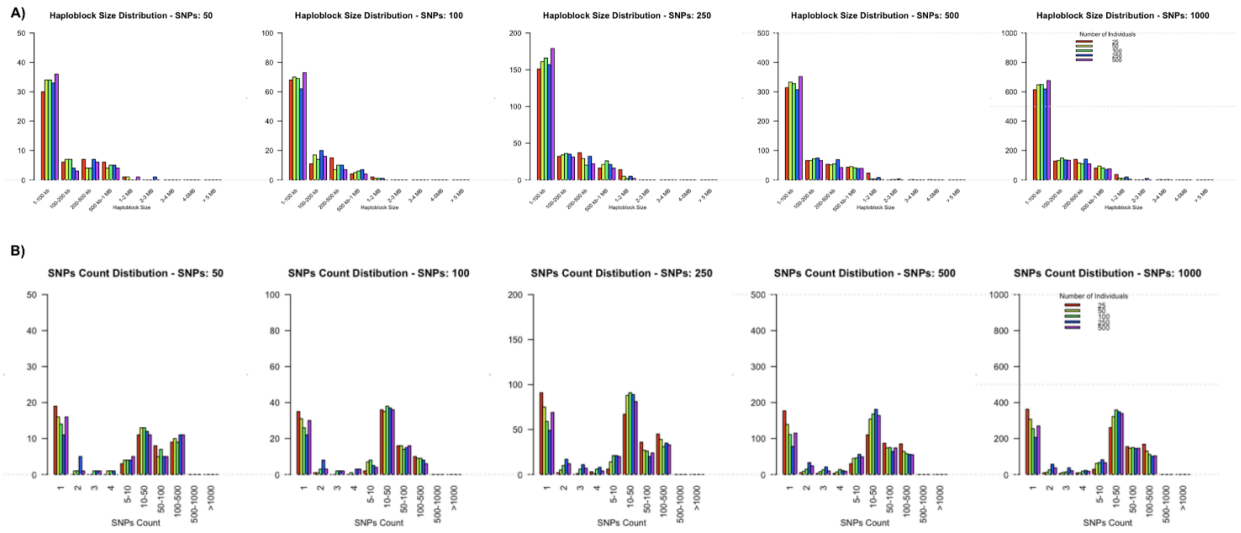

Figure S5 Haploblock Size and SNP Count Distributions for List SNPs Mode Across Different Predefined SNP Lists. *Different bar plots for haploblock size and SNP count distributions generated using the List SNPs mode, with predefined SNP lists containing 50, 100, 250, 500, and 1,000 SNPs, with the option for which we have that the number of the SNPs in the list corresponds to the number of the haploblocks detected. The SNPs of the lists were chosen randomly from a subregion of the chromosome 22 (between positions 16,554,711 bp – 21,512,333 bp) and the input data is from the whole chromosome 22 in genotyped individuals from the DESIR cohort* **(A)** *Haploblock Size Distribution, five separate plots show the distribution of haploblock sizes for each predefined SNP list size. The x-axis displays haploblock size categories (ranging from 1–100 kb to over 5 MB), while the y-axis represents the number of haploblocks observed for each category. Distinct colors are used to represent different sample sizes (25, 50, 100, 250, 500 individuals).* **(B)** *SNP Count Distribution, five plots show the distribution of SNP counts per haploblock for each predefined SNP list size. The x-axis shows SNP count categories (ranging from 1 to >1,000 SNPs per haploblock), and the y-axis represents the number of haploblocks observed. Each sample size is represented by a unique color in the bar plots. These visualizations are based on a smaller genomic region of 5 Mb and were generated using the following parameters: LD thresholds ($r^2 \geq 0.1$, $D' \geq 0.7$), MAF ≥ 0.8, %carrier threshold ≥ 80, region size of 5,000,000 base pairs, and a maximum empty gap of 200 SNPs.*

2.3 Comparison of HaploExplore's results on European and African ancestry datasets

To assess the robustness of HaploExplore, we compared its performance across two independent French cohorts (DESIR and SU.VI.MAX [1]) and two independent datasets of African ancestry (AFR1 and AFR2 [2]). All datasets were analyzed under the standard mode with identical parameters with the same numbers of individuals (N=250) and in the same genomic region (chr22), SU.VI.MAX with 140,634 SNPs, DESIR with 125,956 SNPs AFR1 with 267,251 SNPs, and AFR2 with 273,645 SNPs.

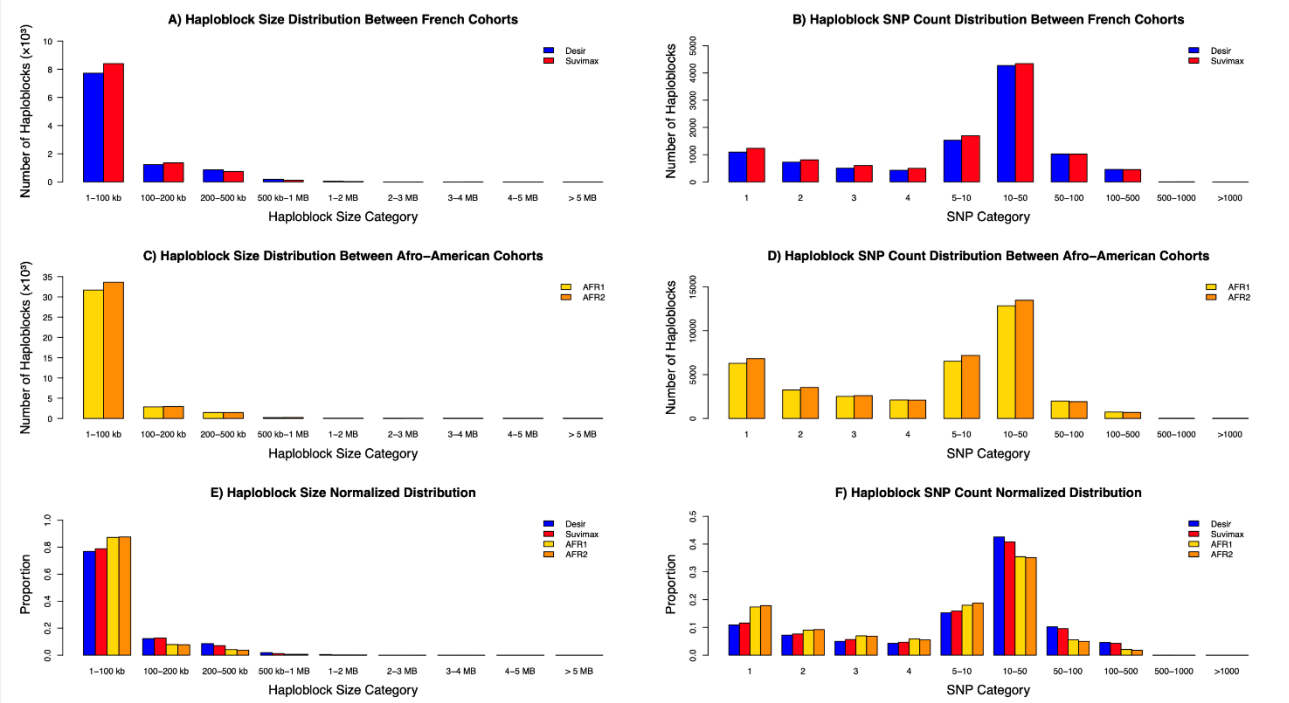

Figure S6. (A) *Haploblock size distribution in two French cohorts (DESIR and SU.VI.MAX). The x-axis represents haploblock size categories, while the y-axis shows the number of haploblocks in each category. Bar plots display DESIR in blue and SU.VI.MAX in red.* **(B)** *SNP count distribution per haploblock in the same two French cohorts. The x-axis represents SNP count categories and the y-axis represents the number of haploblocks observed.* **(C)** *Haploblock size distribution in two Afro-American cohorts (AFR1 and AFR2), with the same axis definitions as in (A).* **(D)** *SNP count distribution per haploblock in AFR1 and AFR2, using the same axis definitions as in (B).* **(E)** *Proportional haploblock size distribution across all four datasets (DESIR in blue, SU.VI.MAX in red, AFR1 in gold, and AFR2 in dark orange).* **(F)** *Proportional SNP count distribution per haploblock across the four datasets. All panels are based on chromosome 22 data using the standard mode of HaploExplore with 250 individuals per cohort. Parameters: LD thresholds ($r^2 \geq 0.1$, $D' \geq 0.7$), MAF ≥ 0.8, carrier percentage threshold ≥ 80%, region size of 10,000,000 base pairs, region overlap of 5,000,000 base pairs, and a maximum empty gap of 200 SNPs.*

The distribution of haploblock sizes between the two French datasets is similar, with the majority of haploblocks falling within the 1-100 kb range (*Figure S6C; S6D*). AFR1 contained 6,297 single-SNP haploblocks, while AFR2 exhibited 6,828. A similar trend was observed across all SNP categories (*Figure 6D*), where AFR2 consistently displayed slightly higher values, reflecting its slightly elevated haploblock count. In both populations, the majority of haploblocks fell within the 1-100 kb range, with 31,691 in AFR1 and 33,620 in AFR2. A chi-square test comparing haploblock size distributions between AFR1 and AFR2 resulted in $\chi^2 = 90$, $p = 0.2313$, indicating non-significant differences in the haploblock structures. Using a five-SNP flexibility, 54.78% of AFR1 haploblocks were also present in AFR2, and 51.85% of AFR2 haploblocks overlapped with AFR1.

Figure S6E and S6F display the histograms of the proportion of each category of haploblocks (size in bp or SNP number) in each population (normalized with the total number of haploblocks in the population). Compared to the French populations (DESIR and SU.VI.MAX), Afro-American

populations (AFR1 and AFR2) displayed a higher number of haploblocks, particularly in the smaller size categories even when we normalized the results for the total number of haploblocks found for each cohort (Statistical analysis using Fisher's Exact Test confirmed that the differences in haploblock distribution between the two groups were highly significant for most size categories; p-values < 2.2e-16 for the smallest haploblock categories). This observation supports the expectation that African populations are characterized by greater levels of genetic diversity and less LD among loci compared to non-African populations [3,4]. The coefficient of variation (CV) for the size, which measures relative dispersion (SD/Mean), is slightly higher in Afro-American datasets (AFR1: 2.58, AFR2: 2.59) than in European ones (DESIR: 2.26, SU.VI.MAX: 2.32), this indicates that, although European populations have larger haploblocks on average, their sizes vary more widely relative to their mean sizes. The higher genetic diversity in Afro-American populations break down long haploblocks over generations, leading to a higher proportion of shorter segments [5]. This explains why Europeans compared to Afro-American exhibit longer, more conserved haploblocks due to genetic bottlenecks and lower recombination rates [5]. The founder bottleneck (~50–100 thousand years ago) in non-African populations reduced genetic diversity, increased LD, and led to more uniform LD patterns due to serial founder effects during migration [5]. These results suggest that HaploExplore demonstrates reliability and flexibility by capturing haploblock structures consistently within homogeneous populations while also reflecting expected population-specific differences.

1. Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D, et al. The SU.VI.MAX Study: A Randomized, Placebo-Controlled Trial of the Health Effects of Antioxidant Vitamins and Minerals. Arch Intern Med. 2004 Nov 22;164(21):2335.
2. McLaren PJ, Coulonges C, Ripke S, van den Berg L, Buchbinder S, Carrington M, et al. Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. PLoS Pathog. 2013;9(7):e1003515
3. Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics. 1964 Jan;49(1):49–67.
4. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005 Oct;437(7063):1299–320.
5. Campbell MC, Tishkoff SA. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annu Rev Genom Hum Genet. 2008 Sep 1;9(1):403–33.

2.4 Comparison of HaploExplore's results on different chromosomes for the DESIR dataset

In addition to the comparison of different populations, we also applied HaploExplore to the full chromosomes 1 and 6 from the French DESIR cohort to assess its robustness. This analysis demonstrated the software's ability to process large chromosomes with a high number of SNPs. It was performed in Standard mode with default parameters on a total of 250 individuals. Chromosome 1 contains 1,633,735 SNPs, chromosome 6 contains 697,665 SNPs, and chromosome 22 contains 125,936 SNPs.

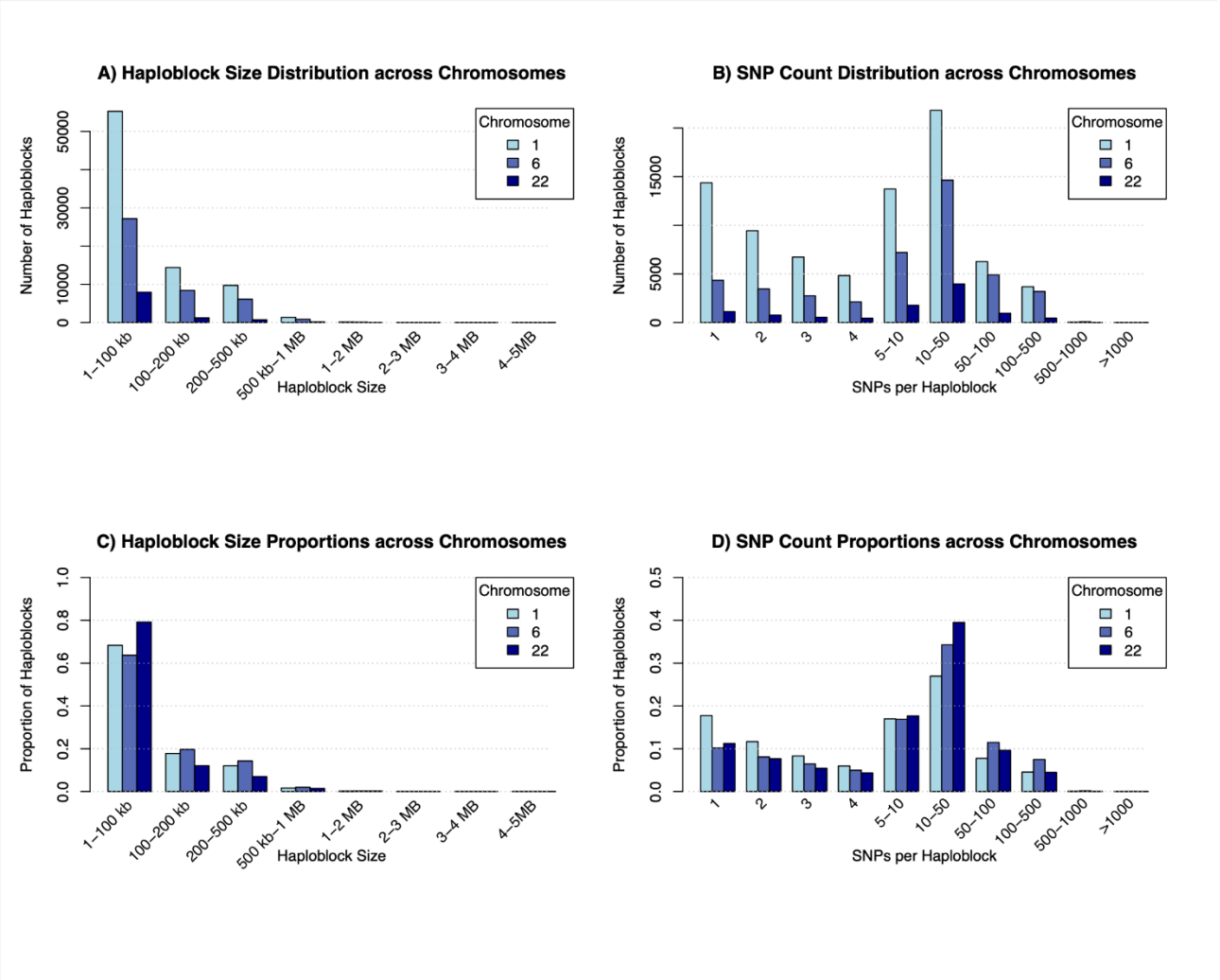

Figure S7. Haploblock size and SNP count distributions across chromosomes 1, 6, and 22. (A) Distribution of haploblocks by physical size (bp) across the three chromosomes. (B) Distribution of haploblocks by SNP count per block across the three chromosomes. (C) Proportional distribution of haploblock sizes across the three chromosomes. (D) Proportional distribution of haploblocks by SNP count across the three chromosomes.

2.5 Results replication

As a benchmark, we evaluated whether HaploExplore was able to correctly retrieve well-characterized haploblocks within the HLA region. These haploblocks had been built by manual computation in a previous study and therefore provide a good test case for validation. With the same cohort, HaploExplore identifies exactly the same haploblocks (sometimes with a difference of 1 in terms of SNP number), confirming the robustness of the approach.

Table S7. Haploblocks identified by manual computations in the HLA region by the study of Rahmouni et al (ref 10 of the the main text) are fully retrieved by HaploExplore.

SNP	Associated MiA-haploblock			
ID	Start	End	Size (bp)	N snps
rs9264942	31268398	31351614	83216	15
rs150908530	29585114	30102498	517384	474
rs1233396	28553793	31228219	2674426	1933
rs79972666	31025467	31512191	486724	71
rs9468885	31174520	31350608	176088	54
rs1894406	32774141	32829520	55379	87

These results were generated from the control group called ill1 of the genotyped International Collaboration on HIV-1 Genomics (ICGH). The haploblocks identified are quasi-identical to the ones computed manually and published in the Table 4 of the publication by Rahmouni et al. (ref 10 of the main text). The position number corresponds to the (GRCh38) version of the human genome.

2.6 Speed

To evaluate their performance further, we tested these modes in three small regions of chromosome 22 from 250 genotyped individuals of the DESIR cohort with varying SNP densities: Region 1 (13,032 SNPs, reduced to 9,303 SNPs after filtering for MAF > 0.01), Region 2 (13,255 SNPs, reduced to 8,824 SNPs after filtering for MAF > 0.01), and Region 3 (19,637 SNPs, reduced to 12,777 SNPs after filtering for MAF > 0.01).

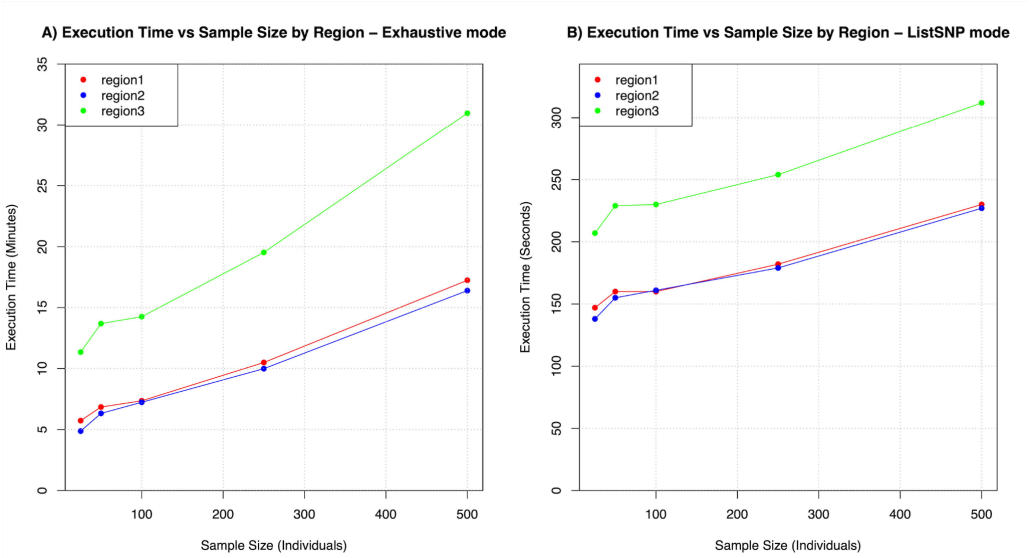

Figure S8 - (A) Execution Time vs Sample Size by Region in Exhaustive Mode, relationship between execution time (in minutes) and sample size (number of individuals) for three different regions (Region 1 : 9303 SNPs, Region 2: 8824 SNPs, and Region 3: 12777 SNPs) using the Exhaustive mode. (B) Execution Time vs Sample Size by Region in ListSNP Mode (SNPs = 1000), relationship between execution time (in seconds) and sample size (number of individuals) for the same three different regions (Region 1, Region 2, and Region 3) using the ListSNP mode with subgroups taken by the DESIR cohort

For the exhaustive mode, the running times were analyzed across these regions with the same sample sizes as before. As expected, the computational time increased with both the number of individuals and the SNP density of the region (Table S8; *Figure 8A*). Region 3, with the highest SNP count, consistently required the most processing time. The *ListSNP* mode's running times were also evaluated across the same regions, with varying sample sizes and fixed SNP list sizes (50, 100, 250, 500 and 1000 SNPs). As anticipated, regions with higher SNP densities, such as Region 3, required slightly more computation time than Region 1 and Region 2. Furthermore, the computational time increased with both the sample size and the size of the predefined SNP list per region (Figure S8B, Table S8). These results are coherent with the ones already obtained for a larger region such as the chromosome 22.

Table S8 Execution Time (in Minutes) for Different Modes with various sample sizes of genotyped individuals from the DESIR cohort (25, 50, 100, 250, and 500 individuals), comparing the Standard, ListSNP, and Exhaustive modes in chromosome 22.

Individuals	Modes	Time (Minutes)
25	Standard	~ 7
50	Standard	~ 11
100	Standard	~ 20
250	Standard	~ 26
500	Standard	~ 88
25	ListSNP	~ 2
50	ListSNP	~ 2

100	ListSNP	~ 2
250	ListSNP	~ 4
500	ListSNP	~ 6
25	Exhaustive	~ 263
50	Exhaustive	~ 352
100	Exhaustive	~ 325
250	Exhaustive	~ 443
500	Exhaustive	~ 1215

To further evaluate computational performance, running times were compared across different imputed chromosomes (chr1, chr6, and chr22) using the Standard mode with default parameters in the DESIR cohort. For a sample of 250 individuals, the analysis of chromosome 1 with 1,633,735 SNPs required approximately 580 minutes. When increasing the sample size to 500 individuals for chromosome 6, which contained 697,685 SNPs, runtime extended to approximately 720 minutes (~12 hours). In contrast, chromosome 22, with 125,956 SNPs, was processed in just 26 minutes for 250 individuals.

Table S9 Running Time (in Minutes) for Exhaustive Mode for three 5 Mb Regions of Chromosome 22 with phased data (Region 1: 9303 SNPs, Region 2: 8824 SNPs, and Region 3: 12777 SNPs), according to the sample size of genotyped individuals from the DESIR cohort

Individuals	Region	Time (Minutes)
25	region1	~ 6
25	region2	~ 5
25	region3	~ 11
50	region1	~ 7
50	region2	~ 6
50	region3	~ 14
100	region1	~ 7
100	region2	~ 7
100	region3	~ 14
250	region1	~ 11
250	region2	~ 10
250	region3	~ 20
500	region1	~ 17
500	region2	~ 16
500	region3	~ 31

In this analysis, we examined the running time of the HaploExplore software in exhaustive mode across three regions of different SNP densities (Region 1: 9303 SNPs, Region 2: 8824 SNPs, and Region 3: 12777 SNPs) for varying sample sizes (25, 50, 100, 250, and 500 individuals).The running time increased with both the number of individuals and the size of the region, as expected

(*Table S9*). As the number of individuals and SNPs in the region increased, so did the computational time. Region 3, with the highest SNP count, consistently required the most time for computation.

Table S10 Average Running Time (in seconds) for ListSNP Mode across three 5 Mb Regions of chromosome 22 (Region 1: 9303 SNPs, Region 2: 8824 SNPs, and Region 3: 12777 SNPs), according to the sample size of genotyped individuals from the DESIR cohort

Region	Individuals	SNPs	Time (Seconds)
region1	25	50	7
region2	25	50	6
region3	25	50	8
region1	50	50	7
region2	50	50	7
region3	50	50	10
region1	100	50	8
region2	100	50	8
region3	100	50	11
region1	250	50	10
region2	250	50	10
region3	250	50	14
region1	500	50	18
region2	500	50	18
region3	500	50	25
region1	25	100	9
region2	25	100	9
region3	25	100	12
region1	50	100	10
region2	50	100	10
region3	50	100	14
region1	100	100	11
region2	100	100	12
region3	100	100	16
region1	250	100	15
region2	250	100	15
region3	250	100	20
region1	500	100	26
region2	500	100	25
region3	500	100	33
region1	25	250	21
region2	25	250	20
region3	25	250	30

region1	50	250	24
region2	50	250	23
region3	50	250	34
region1	100	250	25
region2	100	250	25
region3	100	250	36
region1	250	250	32
region2	250	250	32
region3	250	250	44
region1	500	250	49
region2	500	250	49
region3	500	250	65
region1	25	500	51
region2	25	500	47
region3	25	500	71
region1	50	500	56
region2	50	500	54
region3	50	500	79
region1	100	500	58
region2	100	500	58
region3	100	500	82
region1	250	500	69
region2	250	500	68
region3	250	500	96
region1	500	500	96
region2	500	500	96
region3	500	500	132
region1	25	1000	147
region2	25	1000	138
region3	25	1000	207
region1	50	1000	160
region2	50	1000	155
region3	50	1000	229
region1	100	1000	160
region2	100	1000	161
region3	100	1000	230
region1	250	1000	182
region2	250	1000	179
region3	250	1000	254

region1	500	1000	230
region2	500	1000	227
region3	500	1000	312

The running times for the ListSNP method were measured across three regions (Region 1, Region 2, and Region 3), with varying sample sizes and SNP numbers (Table S10).
As expected, the regions with higher SNP counts, such as Region 3 (12777 SNPs), consistently required more time than Region 1 (9303 SNPs) and Region 2 (8824 SNPs). However, since the number of SNPs was fixed at 50, 100, 250, 500 and 1000 we observe that the running time increases with both the sample size and the SNP list per region.

2.7 Comparison with other haploblock detection software

Tool	Number of Haploblocks	Mean Haploblock Size (bp)	Max Haploblock Size (bp)	Standard Deviation (bp)	Runtime
HaploExplore	547	56,865.19	574,879	67,606.29	*~50 sec*
PLINK	530	5,348.24	284,694	15,057.21	*~2 min*
Big-LD	176	18,032.3	296,637	38,367.54	*~50 sec*
HaploBlocker	143	78,703.35	760,929	94,408.67	*~5 sec*

Table S11 - Haploblock Detection Results for Small Region (Region1 - 13,032 Variants of the chromosome 22 with 250 genotyped individuals from the DESIR cohort), comparison of haploblock detection tools based on the number of identified haploblocks, mean and maximum haploblock size, standard deviation, and runtime.

In table S11, we compared the 4 software, namely HaploExplore, PLINK, Big-LD, and HaploBlocker, on a small region of chromosome 22 (13,032 variants) with 250 random individuals from the DESIR cohort, with their default settings. The results reveal notable differences, likely due to each tool's distinct purpose and haploblock identification method. HaploBlocker, for instance, is designed to represent genetic variation in a compact set of haplotype blocks, leading to fewer but larger blocks. This is reflected in its low haploblock count (143) and high mean haploblock size (~78.7 kb), making it the most broad-scale approach. Big-LD, which detects subgroups of SNPs based on |r| correlation thresholds and partition SNPs using an interval graph, also identified relatively few blocks (176), with intermediate block sizes (~18 kb mean). In contrast, PLINK applied Gabriel's method using D' confidence intervals, leading to smaller and more numerous haploblocks (530 blocks, ~5.3 kb mean size). HaploExplore, designed for MiA-haploblock detection, produced 547 haploblocks, with a mean block size (~56.9 kb) notably larger than PLINK's but smaller than the one of HaploBlocker. Its approach, which orders SNPs by MAF and incorporates LD thresholds and carrier percentage, allows for a flexible and biologically meaningful haploblock definition, particularly suited for minor allele-driven associations. Computational efficiency also varied: PLINK and HaploExplore completed in ~2 minutes, Big-LD in ~50 seconds, and HaploBlocker in just ~5 seconds, highlighting different algorithmic complexities, due to the different haploblock construction (Table 1).

Table S12 Comparison with other software: running time obtained for the haploblock analysis of 250 genotyped individuals from the DESIR cohort in full chromosome 22

Tool	Runtime (chr22)
HaploExplore	~26 min
Plink	~36 min
Big-LD	~7 min
HaploBlocker	~2 min

Table S13 Comparison of haploblocks between HaploExplore and PLINK (Region1 - 13,032 Variants of the chromosome 22 with 250 genotyped individuals from the DESIR cohort)

Summary Statistic	HaploExplore	PLINK
Total Blocks	612	530
Mean SNPs per Block	24.1	16.4
Mean Block Length (bp)	50,827	5,349
Mean SNP Density (SNPs/kb)	0.474	3.066
Median SNPs per Block	12	8
Median Block Length (bp)	26,344	2,318

To show how HaploExplore provides different results compared to other software, we compared haploblocks obtained with our algorithm and with PLINK on a small region of chromosome 22. Our method identified 612 haploblocks, whereas PLINK reported 530. Haploblocks from our algorithm are inclined to include more SNPs (median = 12, mean = 24) than PLINK (median = 8, mean = 16). In terms of genomic region, the haploblocks detected by HaploExplore were longer (median length = 26 kb, mean = 50 kb) compared to those from PLINK (median = 2.3 kb, mean = 5 kb). Overall, the observed differences illustrate how the two approaches diverge in block definition. PLINK defines haploblocks based exclusively on linkage disequilibrium metrics such as D′, which favors the construction of dense haploblocks (3.066 SNPs per kb). In contrast, HaploExplore integrates the carrier percentage parameter, which specifically considers whether minor alleles co-occur with the minor allele of a core SNP. As a consequence, SNPs that do not share the minor allele with the core SNP are excluded from the haploblock, leading to haploblocks that are less dense but extend over larger genomic regions (0.474 SNPs per kb). This allowed HaploExplore to specifically capture MiA-haploblocks, which are overlooked by other methods.