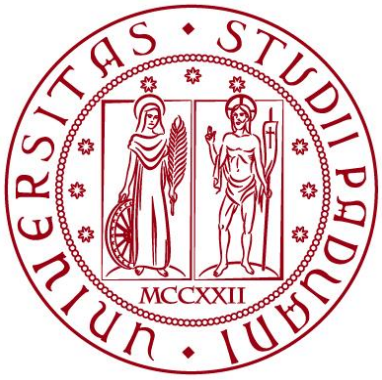
UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea in Biologia



ELABORATO DI LAUREA

RUNS OF HOMOZYGOSITY IN HUMAN POPULATION

Tutor: Prof. Massimo Mezzavilla

Dipartimento di Biologia

Laureanda: Susanna Trevisanello

ANNO ACCADEMICO 2023/2024

INDEX

ABSTRACT

Runs of Homozygosis (ROH) are stretches of haplotypes that are identical by descendance (IBD). The distribution is really various around the globe. ROH load is informative about the history of a certain population, as it is the result of past demographic events. The study of ROH is useful for mapping recessive variants associated with genetic diseases.

In this thesis, the main aspects regarding ROH are shortly exposed by referring to some articles found in literature about the topic. To apply practically what was previously learned, simple scans for different categories of ROH, were run using the software Plink and the statistical analysis were done with Excel. The aim was to test that the genome of different human groups is loaded with different ROH in length and number according to the demographic history and cultural costumes.

ITALIAN:
Le "Runs of Homozigosity" (ROH) sono tratti di aplotipi definiti identici per discendenza. La distribuzione varia tra le diverse popolazioni umane del mondo. Il carico di ROH da informazioni sulla storia di una certa popolazione, in quanto è il risultato di eventi demografici passati. Inoltre, lo studio delle ROH è utile per la mappatura delle varianti recessive legate a malattie genetiche.

In questa tesi, vengono esposti gli aspetti principali riguardanti le ROH facendo riferimento ad alcuni articoli della letteratura scientifica sull'argomento.

Per applicare nella pratica, quanto appreso in precedenza, sono state eseguite semplici scansioni per diverse categorie di ROH, utilizzando il software Plink e l'analisi statistica è stata effettuata con Excel. L'obiettivo era di verificare come il numero e la lunghezza delle ROH varia tra gruppi di individui appartenenti a diverse aree geografiche, sulla base degli eventi demografici passati.

KEYWORDS:
IBD, admixture, bottleneck, consanguinity, demographic history

1. INTRODUCTION

Runs of Homozygosity (ROH), also called by some authors Runs of Autozygosity (ROA), are long stretches of DNA in a homozygous state. They manifest when an individual inherits the same haplotype, defined as “identical by descendance” (IDB), from both parents as the result of recent relatedness. Runs of Homozygosity are ubiquitous in the genome and in the human population but the scenario is very varied and complex.

The study of the length, the number and distribution of ROH in worldwide human populations offers an insight into demographic dynamics and allows the reconstruction of population history as the ROH load is shaped by specific demographic events.

The increase in homozygosity is linked to enrichment in deleterious alleles, therefore the study of ROH can be useful to explore genetic diseases linked to loss-of-function variants.

1.1 Origin of ROH and demographic processes

“RHO arise when two copies of an ancestral haplotype are brought together in an individual: longer haplotypes inherited from recent common ancestors or shorter haplotypes from distant ones. “(Ceballos et al.,2018)
In other words, two related individuals share part of their genome and their offspring are likely to inherit identical segments of chromosomes. The more closely related they are, the longer the shared part of the genome is, while the more distant they are, more differences are accumulated in their genome. The stretches of DNA identical by descendance decrease with increasing time and effective population size, in fact recombination had more “occasions” to break down haplotypes.

Usually, ROH are divided into three classes. Short ROH are tens of kilobases long, they are old haplotypes and reflect distant relatedness; medium ROH are hundreds of kilobases long and reflect background relatedness in the population; finally long ROH are from hundreds of kilobases to several mega bases long and are the result of recent parental relatedness. (Szpiech et al.,2017)

As we humans are all closely related to some degree, ROH are very common and universally present in the human genome, even in outbred populations.

Populations are characterized by differences in number (NROH) and length (SROH) of ROH, depending on the processes that each population was involved in.

It is important to consider the effective population size when speaking of ROH. Small populations are characterized by a higher NROH and SROH than large populations.
Admixed populations have the lowest NROH and SROH. Admixture is the phenomenon where two separated populations mix. As the two populations are distantly related and very highly differentiated between each other, identical stretches of DNA are very reduced. Therefore, the quantity of ROH that arise in the offspring is low.
On the other hand, in populations characterized by high levels of consanguinity (mating between cousins), all individuals are closely related. Consequently, ROH increase mainly in length but also in number.
Bottleneck is a random event that reduces the size of a population and its genetic variability. It reduces by chance the number of ancestral lineages in a population in comparison to the original one. As a consequence, all individuals off the bottlenecked population derive from a small number of founders, increasing consequently NROH and SROH.
The history of populations with higher NROH and SROH are characterized by the combination of the two phenomena: consanguinity and bottleneck.

It is very interesting to see how genetics is strongly linked to history, demography and traditions. In some cultures, endogamy (marriage within the same group or community) or consanguinity are encouraged, therefore cultural traditions and costumes play a central role together with migration events, isolation and size of the populations to shape differently in each population the distribution of ROH. That is why the study of Runs of Homozygosity might be helpful to reconstruct some population dynamics and opens a window to explore evolution and genetics.

1.2 ROH in different populations

Ceballos et al. (2018) in their article divided human populations in five classes analyzing long and short ROH.

The first class consists of some consanguineous populations: where a high mean of SROH is measured due to recent inbreeding. Some examples are many Muslim communities in Dagestan, Pakistan and West Asia.

The second class, characterized by individuals enriched in short ROH and poor in long ones. This class includes populations, such as Papua New Guinean Highlanders, Koryak and Chukchi in Siberia and Athabaskans in North America, the result of isolation and endogamy a long time ago and no recent inbreeding, therefore there is no recent common ancestry.

A third class is composed of some Native American populations, where both ancient and recent inbreeding took place, therefore individuals are enriched for both long and short ROH.

The most numerous in the globe is the fourth class: populations with large effective population size where ROH are reduced in number and length.

The fifth class is composed of admixed populations such as Latinos or African Americans. The picture in this case is very varied depending on the specific history and ancestry component of each group.

1.3 Admixed populations: ancestry dependent enrichment of ROH

It is very interesting to analyze the complex situation of admixed populations, seeing how differences in NROH and SROH depend on the percentage of admixture and on the differences between parental haplotypes.

Szpeich et al. (2019) in their article, starting from three admixed populations: Mexican Americans, African Americans and Puerto Ricans, with different European, Native American and African contribution, analyzed their ancestral component, quantified ROH and the accumulation of deleterious variants in ROH.

Puerto Ricans have mostly African and European ancestry, Mexican Americans have European and Native Americans ancestry, while African Americans have mostly African and European ancestry.
In each parental population, total length of ROH is directly proportional to the distance from Africa. In fact, Europeans and Native American have more ROH compared to Africans as a consequence of the various bottleneck events during the dispersion around the globe.

Mexican Americans have the highest total number of ROH followed by Puerto Ricans and African Americans and except for the short ROH, they have the highest total length of medium and long ROH. These data can be explained by the fact that the Mexican population, derives from admixture

between two parental populations, distant from Africa, that suffered many bottleneck events that increased NROH and SROH. On the other hand, the other two admixed populations, have an important component of African ancestry which is poorer in ROH, particularly in long and medium ones.

1.4 Distribution of ROH in the genome

ROH are ubiquitous and frequent in the human genome, but they are not equally distributed across it.

Regions characterized by low recombination and therefore by low genetic diversity are rich in ROH. Recombination favors reshuffling by breaking ancestral haplotypes and creating new combinations of alleles. It means that genetic diversity increases and the ancestral haplotype disappears rapidly decreasing the probability of being autozygous. In regions of the genome where recombination is low, the original haplotype, deriving from a common ancestor, is kept and it is much more likely that it “meets” with an identical one.
An important example is the X chromosome, characterized by low recombination rate, except for the pseudo autosomal regions. Being for one third of its time in the male germline, it cannot recombine, therefore it is characterized by low diversity.

Ceballo et al. (2018) speak of “ROH islands", referring to regions of the genome with a high percentage of ROH. They are very common in all populations and are dominant in outbred individuals, where the majority of ROH are found in the islands and just a few short ROH are found outside.
On the other hand, in recently inbred individuals these islands are overshadowed by longer ROH spread randomly around the genome. This happens because time was not enough for recombination to break down identical segments and therefore, they are randomly distributed in the genome instead of being concentrated in specific regions, characterized by low recombination rate.

Szpiech et al. (2017) created two gene sets to verify if some genes are more enriched in than others. The first set, named high-pLI, consists of genes that are likely to be more intolerant to loss of function, therefore mutations may be more deleterious on average. The second set was called low-pLI, characterized by genes more tolerant to loss of function.
They found that high-pLI genes have, on average, more ROH. This is due to the fact that these genes experience a stronger background selection,

where there is a loss of diversity due to elimination of deleterious alleles and at the same time accumulation of homozygous regions.

1.5 Enrichment in deleterious variants and disease risk

Early studies have identified an association between inbreeding and increased risks of artery diseases, stroke, cancer and neurodegenerative diseases such as Parkinson and Alzheimer.
The alleles associated with these pathologies are usually recessive and not very common. With inbreeding or bottleneck events, these variants are likely to become more common and in homozygosis.
These patterns highlight how cultural and population processes shaping ROH levels in the genome can increase the level of homozygotes in the population, with possible negative consequences for the health of the individuals.

Under generalization, it is possible to observe that the rate of accumulation of deleterious variants in long ROH is higher than that of beneficial variants. While in short ROH we observe the opposite situation.
Consequently, long ROH that are recent haplotypes, compared to short ones usually carry more deleterious alleles. This happens because purifying selection had more time to act in older haplotypes by eliminating strongly deleterious variants.
Moreover, the gain of strongly damaging homozygotes is higher than the gain of moderately and weakly damaging ones and they will tend to be concentrated in autozygous regions, highlighting the correlation between ROH and increased risk disease.

The scenario is very complex and various, due to variety in populations, backgrounds, traits and diseases therefore increase in ROH cannot be always used to indicate increased number of deleterious variants and increased disease risk.
That why the understanding of ROH and the enrichment in deleterious variants, can help track and study some Mendelian and non-Mendelian genetic diseases in the human worldwide population.

2. PRACTICAL APPLICATION

Some genomic data have been analyzed in order to quantify and compare the distribution of ROH around the globe.
The parameters that have been measured are the number of segments (NSEG) and the total length covered by ROH in the genome of an individual (KB). The dataset that was used, included genomes of individuals from all around the world. They were divided into macro groups according to their geographical origin. The areas are Africa, America, Central South Asia, East Asia, Europe, Middle East and Oceania.
The division in geographical areas allows to make considerations about the demographic history of each group that have shaped the load of ROH differently.
Different categories of ROH have been tested, in order to better identify phenomena like recent inbreeding and remove the “background noise” of short ROH that are the result of ancient relatedness.

2.1 Materials and methods

The source of the data used for the analysis is “Insights into human genetic variation and population history from 929 diverse genomes” (Bergström et al., 2020).
The aim of Bergström’s work was to better study and understand the genetic variation in the human species around the globe and how it was shaped by separations of past populations, admixture, adaptation, changes in size and gene flow from ancient human groups. Before this study, large-scale genome sequencing was limited to large metropolitan populations or to small numbers of individuals per group.
The group of scientists sequenced 929 genomes from 54 human populations, using DNA extracted from lymphoblastoid cell lines and incorporated data from a subset of samples that had been previously sequenced. They used the data to measure the variation between and within macro groups. As the genome sequences were made public and freely available by the authors, the data set was used for the analysis of ROH in this thesis.

In order to analyze the dataset, a free software named Plink was used.
Plink is widely used for whole-genome association studies (GWAS) and more generally, in population genetics as it allows the analysis of genomic data.

It is possible to run a simple screen for ROH for any individual using simple commands (--homozyg is the most basic and easy one) and the software generates files with information about the number of homozygous segments, the total coverage of the homozygous portion of DNA, SNP at the start and the end region, physical position, etc. for each individual in the dataset.

It is also possible to "filter" what is identified as a ROH by defying a minimum threshold length value. By default, only Runs of Homozygosity of total length ≥ 1000 kilobases, are noted. It is possible to change these minimums with--homozyg-kb, specifying the length in kb of ROH of interest.

The scan was run first with the command –homozyg and then with –homozyg-kb, respectively for 5000kb, 10000kb and 20000kb. Each step, just longer ROH are considered while the shorter are not counted anymore. The comparison between the scans with different filters allows to make consideration about inbreeding and bottleneck events and to distinguish recent than later relatedness between and within groups.

In this case the parameters of interest were NSEG and total KB, so among the files generated by Plink the file of interest containing this information was “hom.indiv” for each scan.
Subsequently Excel was used for the statistical and graphical analysis.

2.2 Results and discussions

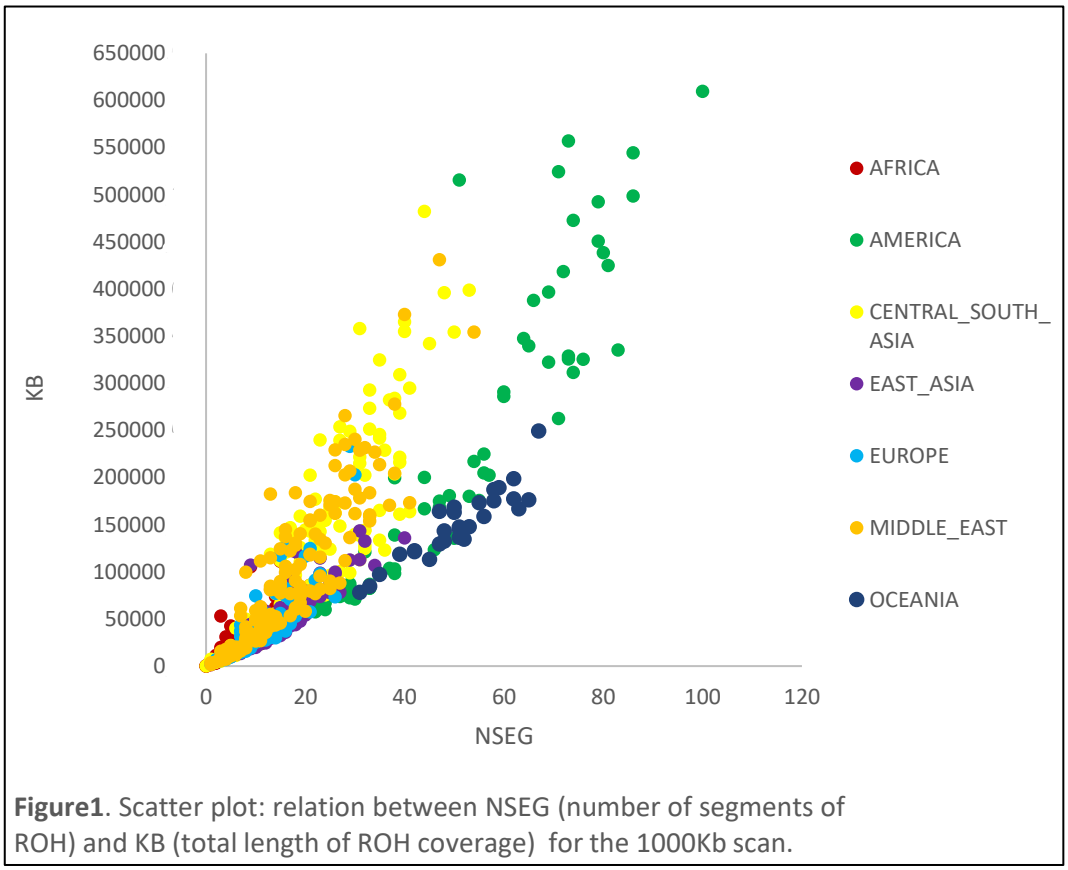


Figure1. Scatter plot: relation between NSEG (number of segments of ROH) and KB (total length of ROH coverage) for the 1000Kb scan.

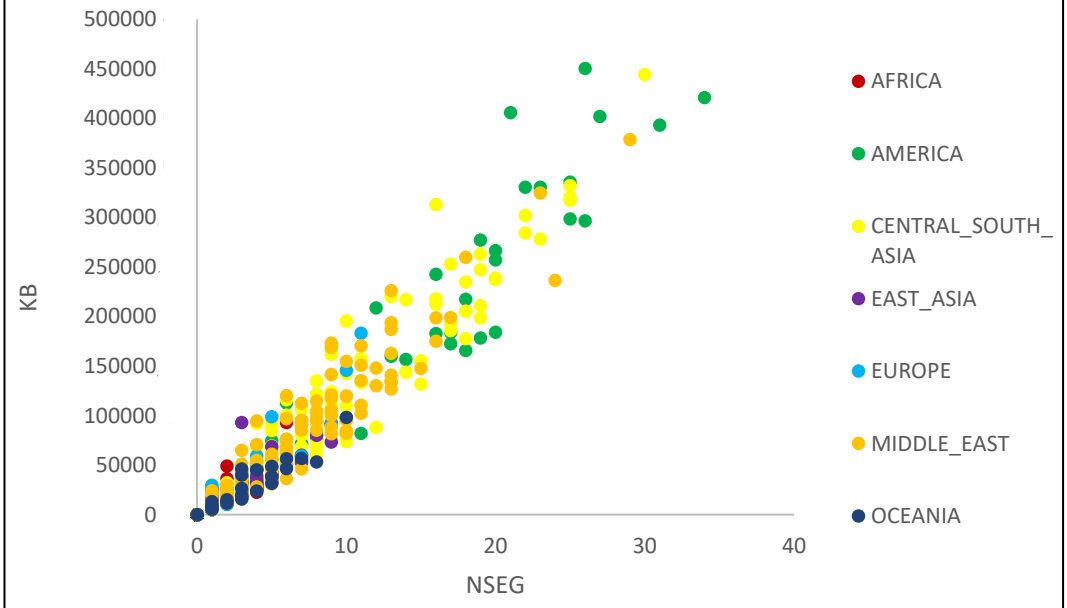


Figure2. Scatter plot: relation between NSEG (number of segments of ROH) and KB (total length of ROH coverage) for the 5000Kb scan.

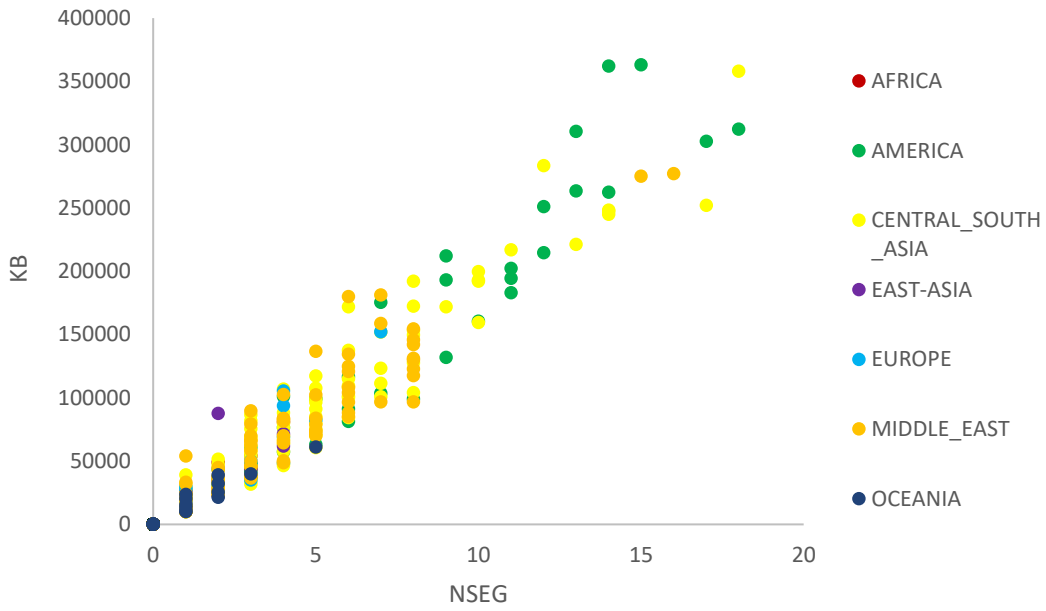


Figure3. Scatter plot: relation between NSEG (number of segments of ROH) and KB (total length of ROH coverage) for the 10000Kb scan.

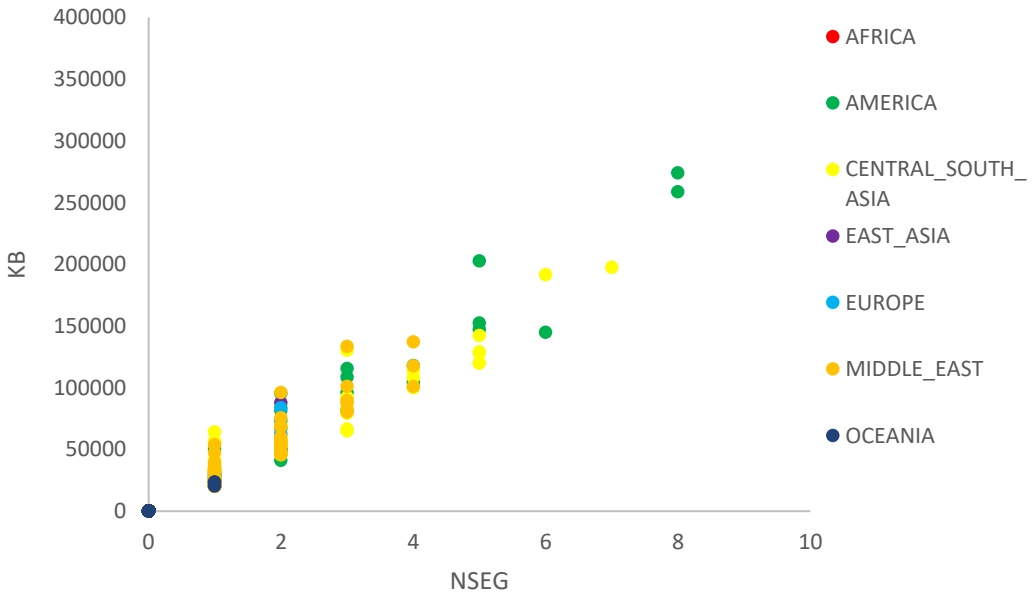


Figure 4. Scatter plot: relation between NSEG (number of segments of ROH) and KB (total length of ROH coverage) for the 20000Kb scan.

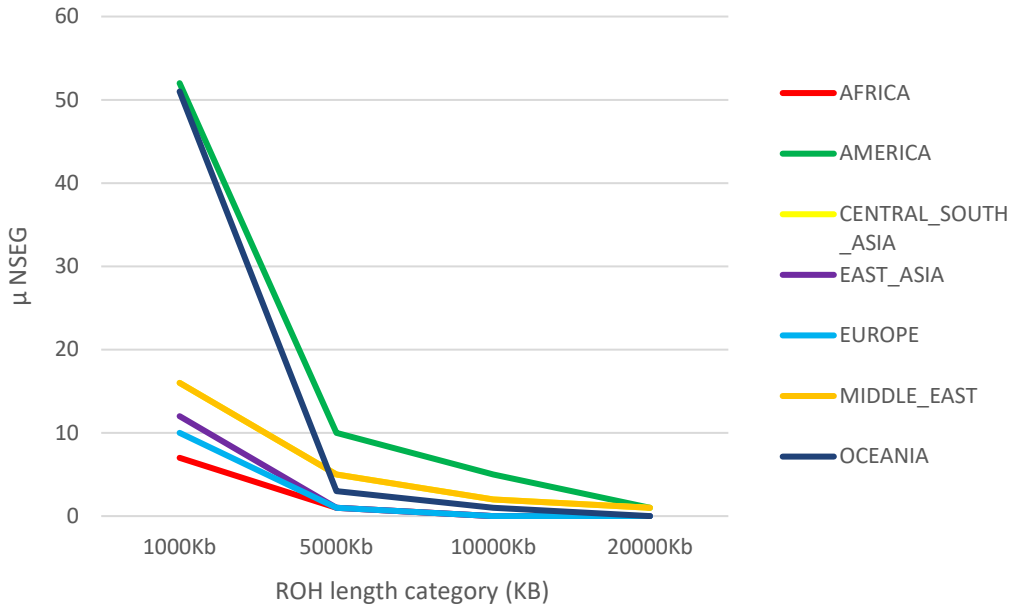


Figure 5. 2-D line graph: µNSEG trend for each geographic area in relation to ROH lenght category in Kb **.** The line for Central South Asia and Middle East appear both in orange, as they overlap.

Table 1 average and standard deviation of NSEG (number of segments of ROH) and KB (total coverage of ROH in the genome) for each geographical area for the 1000Kb scan.

	µ NSEG	σNSEG	µKB	σKB
AFRICA	7	5	27882	26350
AMERICA	52	21	237543	162235
CENTRAL_SOUTH_ASIA	16	12	93276	99308
EAST_ASIA	12	6	37896	26394
EUROPE	10	5	32984	31573
MIDDLE_EAST	16	10	87996	79900
OCEANIA	51	9.5	149908	36612

Table 2 average and standard deviation of NSEG (number of segments of ROH) and KB (total coverage of ROH in the genome) for each geographical area for the 5000Kb scan.

	μ NSEG	σNSEG	μKB	σKB
AFRICA	1	2	12318	19171
AMERICA	10	10	127718	133687
CENTRAL_SOUTH_ASIA	5	6	62502	85780
EAST_ASIA	1	2	9115	16455
EUROPE	1	2	10428	25191
MIDDLE_EAST	5	5	58048	68593
OCEANIA	3	2	27537	22358

Table 3 average and standard deviation of NSEG (number of segments of ROH) and KB (total coverage of ROH in the genome) for each geographical area for the 10000Kb scan.

	μ NSEG	σNSEG	μKB	σKB
AFRICA	0	1	5364	12849
AMERICA	5	5	89815	106185
CENTRAL_SOUTH_ASIA	2	4	42853	66638
EAST_ASIA	0	1	3718	10976
EUROPE	0	1	5904	19590
MIDDLE_EAST	2	3	39039	52259
OCEANIA	1	1	11122	16213

Table 4 average and standard deviation of NSEG (number of segments of ROH) and KB (total coverage of ROH in the genome) for each geographical area for the 20000Kb scan.

	μ NSEG	σNSEG	μKB	σKB
AFRICA	0	0	1675	7258
AMERICA	1	2	41592	64932
CENTRAL_SOUTH_ASIA	1	2	19222	36686
EAST_ASIA	0	0	1226	7272
EUROPE	0	0	2539	11304
MIDDLE_EAST	1	1	17449	29039
OCEANIA	0	0	2407	6981

Generally speaking, by looking at figure 1, 2,3 and 4, it is possible to notice that there is a linear relationship between the number of segments and the total length of ROH in the genome. In other words, by increasing the number of segments, also the total length of ROH per individual increases. Obviously, the trend is slightly different in each group and for each different class of ROH that was tested.

As we could expect, before looking at the results, by increasing the minimum value of what is consider a ROH by the software Plink, the number of NSEG and KB decreases. Observing the figure 5 it is possible to confirm this expectation.

Again, the trend is slightly different in each geographical group depending on the load of long and short ROH and consequently the curve is more or less sharp.

As previously said, ROH load is correlated to the distance of a group of individuals from the African ancestors. Populations distant from Africa are rich both in long ad short ROH. Long ROH are the result of many bottleneck events that they experienced during the dispersion around the globe, while short ROH are the “legacy” from the African ancestry. As consequence in some populations, like Native Americans, the total number of ROH and the total coverage in the genome will be higher than in Africans, whose genome is characterized just by short ROH. Africans also have fewer and shorter ROH because of their larger effective population size.

Observing all the graphs and the tables, is possible to notice that African individuals on average have the lowest number and total length for each ROH category that was tested. Looking at the tables in order (from 1 to 4), it is possible to notice how the mean for NSEG and KB decreases. This confirms the fact that Africans have just a few short ROH.

The American individuals whose genome was sampled, belong to isolated native populations of South and Central America. The origin of the individuals that were tested is represented in figure 6. These communities were not affected by the latest events of migration and admixture that for example involved other places of America or Eurasia.

The results obtained, show that Americans have high ROH load, that includes both long and short ones.

This result is consistent with the idea that the groups that were sampled, created their own “ecosystem” isolated from the rest of the world. Individuals of these communities are likely to be all closely related, due to the isolation from other American populations, that for example experience admixture and gene flow. Furthermore, these communities faced many bottlenecks

events, being very far away from Africa and really isolated. All these factors must have contributed in the enrichment for both long and short ROH.

Figure 7 (Ceballos et al., 2018) shows that Native Americans have the highest total coverage of ROH in the genome, consistently with what it was found in this analysis. All these considerations are also consistent with the results obtained by Bergström et al. in their study: American individuals are rich in private alleles, that were identified and studied. It reinforces again the idea that isolation of a certain group of individuals lead to a independent evolutionary history.

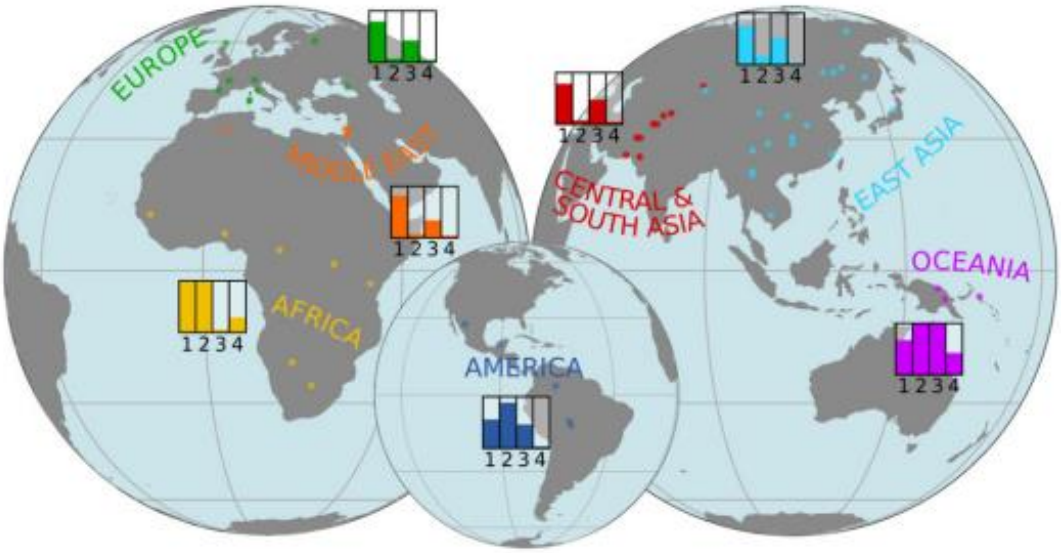


Figure 6. Bergström et al. (2020). Geographical regions of the origin of the individuals whose genome was sampled.

Pacific islanders have a higher burden only for short ROH (Ceballos ed al.,2018). This affirmation reflects the statistics presented above. In figure 1 and table 1 the average number of segments and the total length of ROH per individual for Oceania is high compared to the other groups. Indeed, it nearly reaches the level of Americans, having one of the highest loads of ROH among the groups tested.

The fact that the average for NSEG and KB decreases really fast when scanning for ROH, longer than 5000Kb, confirms the idea that Oceanians are loaded just with long ROH. Short ROH represent ancient relatedness. The absence of long suggests the absence of recent events of inbreeding and bottleneck.

In figure 5, it is noticeable how the average of NSEG in relation to total coverage of ROH for Oceania decays fast, for example comparing to America, whose curve is less sharp, even if the “starting” point for the 1000Kb class is similar.

Eurasia was characterized by events of migration and admixture in the past 10000 years, that for example did not involve the isolated regions of America that were tested. Admixture and gene flow are known for reducing homozygosity and increasing diversity. Furthermore, European and Asian populations are not as far away from Africa as the American population is, so the dispersion was “faster” and less bottleneck events were needed. As consequence on average Europeans and Asians have fewer ROH, both long and short. This affirmation can be confirmed looking at the statistics obtained.

Speaking of Central South Asia and the Middle East, the load of long ROH is consistently high, comparing it for example to Europe and East Asia. It is probably due to recent inbreeding and also to small population size. In many populations of South and Weast Asia consanguinity and endogamy is really common, because it is a cultural practice, but also because the population size is often small. (Ceballos et al., 2018)

The high value of the standard deviation for the NSEG and the KB (table 1, table 2, table 3, table 4) within the different groups, highlights the big variability that characterizes the human species, not just between individuals living in different geographic areas but also within the same “geographical” group.

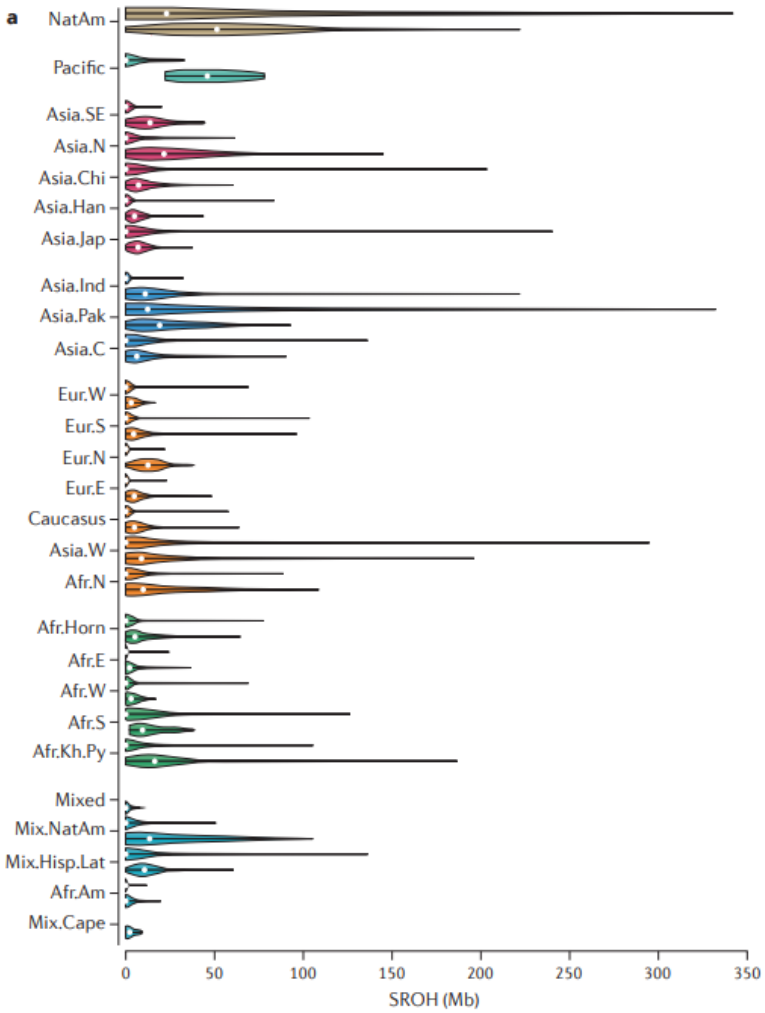


Figure 7 Ceballos et al. (2018). Global census of ROH.

3. CONCLUSIONS

Runs of Homozygosity arise when the same haplotype is inherited by both parents. As recombination is a mechanism that breaks down haplotypes and increases diversity, “the more recent the genealogical relationship of the two parents is, the more frequent and longer the resulting ROH tends to be” (Ringbauer et al., 2021).

ROH are ubiquitous and frequent also in outbred populations but they are not distributed evenly in genome nor in different populations of the world. The different past history of each human group shaped uniquely the load of ROH. Many bottleneck events that allowed the dispersion of our species in all continents starting from Africa, the numerous migrations that were source of gene flow and admixture, together with some cultural practices or isolated distribution that favored inbreeding gave different contribution to the ROH load of each population.

As it was demonstrated in this short thesis with simple analysis, it is possible to identify differences in load of ROH of individuals grouped by their geographical origin. Not just the number, but also the length of ROH changes, reflecting the demographic history, but also the cultural costumes of groups of individuals. It is important to highlight that the scenario is various not just between but also within populations, therefore it is much more complex and deep than how it was presented and analyzed in this short work.

In conclusion, the study of ROH has many potential applications and opens a window that goes beyond the mechanical analysis of genome data. The distribution of human ROH has provided insights into evolutionary, population and medical genetics. By examining the genomic prevalence in a population, we can make discoveries about history and adaptation of human populations and we can learn about the genetics of complex phenotypes. (Szpieich et al., 2019).

In medicine, the study of ROH enriched with deleterious variants it is key for the understanding of Mendelian and more complex genetic diseases (Pemberton et al., 2018).

4. BIBLIOGRAPHY

- Bergström A., McCarthy S.A., Hui R., Almarri M.A., Ayub Q., Danecek P., Chen Y., Felkel S., Hallast P., Kamm J., Blanché H.,Deleuze J-F., Cann H., Mallick S., Reich D., Sandhu M.S. Skoglund P., Scally A., Xue Y., Durbin R. and Tyler-Smith C. (2020) Insights into human genetic variation and population history from 929 diverse genomes . Science 367(6484).

-Ceballos F.C., Joshi P.J., Clark D.W., Ramsay M. and Wilson J.F. (2018) Runs of homozygosity: windows into population history and trait architecture. Nature 19: 220-234.

- Pemberton T.J. and Szpieich Z.A. (2018) Relationship between Deleterious Variation, Genomic Autozigosity, and Disease Risk: Insights from The 1000 Genome Project. The American Journal of Human Genetics 102: 658-675.

- Ringbauer, H., Novembre, J. & Steinrücken, M. (2021) Parental relatedness through time revealed by runs of homozygosity in ancient DNA. Nat Commun 12, 5425

- Szpieich Z.A., Blant A. and Pemberton T.J. (2017). GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification. Bioinformatics 33 (13): 2059-2062.

-Szpiech Z.A., Mak A.C.Y., White M.J., Hu D., Eng C., Burchard E.G., Hernandez R.D. (2019) Ancestry-Dependent Enrichment of Deletrious Homozygotes in Runs of Homozygosity. The American Journal of Human Genetics 105: 747-762.

-Plink version 1.9. Technical documentation. https://www.cog-genomics.org/plink/.