Investigating the Rate of Heterozygosity and Breast Cancer Among Global Populations

Lauren Ji[1] and Sofia Floody[2]

[1]Johns Hopkins Department of Computer Science and [2]Johns Hopkins Department of Computer Science

[1]lji13@jhu.edu[2]sfloody1@jhu.edu

ABSTRACT

The goal of our project is to investigate the relationship between heterozygosity in the BRCA1 and BRCA2 genes and breast cancer incidence and mortality rates across global populations. Breast cancer remains one of the most prevalent cancers affecting women worldwide, with the BRCA1 (Breast Cancer Susceptibility Gene 1) gene and the BRCA2 (Breast Cancer Susceptibility Gene 2) playing critical roles in hereditary breast cancer susceptibility. Using genomic data from The International Genome Sample Resource, we analyzed BRCA1 and BRCA2 sequences from individuals across diverse geographic regions and identified heterozygous single nucleotide polymorphisms (SNPs). We aggregated heterozygosity data by geographic region and compared these metrics with breast cancer incidence and mortality rates obtained from the International Agency for Research on Cancer. Our analysis aims to determine whether regional variations in heterozygosity of the genes in question correlate with breast cancer occurrence patterns globally. This research contributes to understanding the genetics regarding breast cancer susceptibility across populations and may inform customized treatment in different geographic regions.

KEYWORDS: Heterozygosity, Breast Cancer, Incidence, Mortality

INTRODUCTION

Breast cancer is among the most common types of cancers in female populations, impacting millions of women worldwide. The incidence and mortality rates of breast cancer vary considerably across populations, with genetic, environmental, and socioeconomic factors all contributing to these disparities. Understanding the genetics behind breast cancer susceptibility is crucial to developing screening programs and targeted interventions.

The BRCA1 and BRCA2 genes have been established as tumor suppressor genes associated with hereditary breast and ovarian cancer syndromes [9]. Inherited gene mutations in BRCA1, BRCA2, and PALB-2 significantly increase breast cancer risk. Women who have mutations in these genes may benefit from preventive measures such as surgical removal of both breasts and chemoprevention [10]. While pathogenic mutations in BRCA1 and BRCA2 have been studied, the role of genetic heterozygosity in breast cancer susceptibility remains less understood, particularly across global populations.

In countries with a high Human Development Index (HDI), 1 in 12 women are diagnosed with breast cancer in their lifetime, and 1 in 71 women die of it. In contrast, in countries with a low HDI, only 1 in 27 women is diagnosed with breast cancer in their lifetime, while 1 in 48 women will die from it [9]. By studying the relationship between heterozygosity in the BRCA1 and BRCA2 genes and breast cancer incidence and mortality across global populations, we hope to inform the development of population-specific risk assessment models and screening strategies, contributing to more equitable and effective breast cancer prevention efforts worldwide.

Previous data has shown that higher heterozygosity rates can correlate to the increased presence of generally desired traits, such as height and healthy aging [1, 2]. Since breast cancer can be a hereditary trait, we would expect higher heterozygosity rates to correlate to lower rates of breast cancer within a population. Due to previous studies on global heterozygosity rates, we would expect to see that Africans have the highest heterozygosity rates, followed by Americans, South Asians, Europeans, then East Asians.

We would expect that breast cancer rates follow this same pattern, with lowest occurrences in Africans, then Americans, South Asians, Europeans, and East Asians.

Our hypothesis is based on previous data regarding the effects of heterozygosity on traits unrelated to cancer. We are basing our hypothesis on the assumption that cancer is largely based on genetic factors. We acknowledge that cancer is a complicated disease that is affected by many factors, some completely unrelated to the genome. However, to form a hypothesis that fits the scope of this project, we will temporarily disregard other factors.

We aim to determine if there is a correlation between heterozygosity and the incidence and mortality of breast cancer on an individual and global scale. We will do this by counting heterozygous SNPs in the BRCA1 and BRCA2 genes, putting this value into the context of the genome, and then comparing these values with the incidence and mortality rates of the countries from which each sample originated from.

The genomic data used was taken from the 1000 Genomes Project. We decided to exclusively use sequences from female subjects because 99% of breast cancer incidences are in female patients [10]. The data used was from a variety of global populations, with representations from Asia, Africa, Latin America, and Europe.

We obtained data regarding the incidence rate, mortality rate, and HDI values for breast cancer by country from the International Agency on Cancer Research (IARC). We used this data to compare heterozygosity globally.

Our results showed an overall negative correlation between genetic heterozygosity and breast cancer rates across global populations.

METHODS

Part 1

Upon beginning this project, we decided to investigate solely the BRCA1 gene. BRCA1 resides on chromosome 17. We initially used a reference sequence from the Genome Reference Consortium Human Build 38 (GRCh38 (hg38)). We selected this reference because it is very commonly used as a reference sequence, used by databases such as the Cancer Genome Atlas (TCGA). We used The International Genome Sample Resource to obtain sequences in the form of FASTA files from the BRCA1 gene from female individuals from different places around the world [7].

Tools

We aligned our sequences to the reference sequence using the alignment tool, Bowtie2. We used the tool Samtools to compress and standardize our sequences. We then found SNPs using the tool FreeBayes. We performed variant calling and file manipulation to filter out insignificant SNPs and to highlight heterozygous SNPs using BCFtools. We then compressed these files and used BCFtools to index them. Our final files contained heterozygous SNPs within the BRCA1 regions. We aggregated our data of heterozygous SNPs by region of the world, computing the average number of heterozygous SNPs per region.

Analysis

To analyze this data, we compared these values to regional breast cancer incidence and mortality rates from the IARC [8]. We created scatterplots, with a number of heterozygous SNPs on the x-axis and incidence/mortality on the y-axis. The first set of scatterplots compared heterozygosity with incidence. One graph included each sample, and the following graph averaged the values per country and graphed these averages. The second set of scatterplots followed the same pattern, but with mortality instead of incidence. Finally, we computed correlation coefficients to determine if there is a correlation between heterozygous SNPs and incidence on the individual level, incidence on a country level, mortality on the individual level, and mortality on a country level.

Problems from Part 1

A problem we ran into in our initial pipeline was that the process of downloading the full genome files, as well as the alignment and sorting of the files were extremely costly in terms of both memory and time. The process was extremely inefficient and required real-time deletion of files, as our computers' storage filled in the process and consequentially halted before all the analysis was complete, complicating the process.

Next Steps

The results we yielded from our first pipeline were quite weak and inconclusive. As a result, we decided to expand our project. We decided to include the BRCA2 gene on chromosome 13 in our analysis, as well as adding controls and increasing the number of sequences we processed. The control regions are intervals on chromosomes 17 and 13 that are matched in size to BRCA1 and BRCA2 genes but are not part of any known genes.

Part 2

The next step of our project evaluates heterozygous SNPs from the BRCA2 gene and two control regions. When deciding to pivot from our original project, we were aware that this would mean adjusting and rerunning our original pipeline. We were hesitant to rerun the project, since it was so time consuming to run originally. To mitigate this issue, we downloaded pre-processed VCF files that contained heterozygous SNPs already listed in the file.

Tools and New Data

In the process of learning how to access VCF files, we learned that there is a large abundance of samples that

had been processed that were not aligned to GRCh38, but rather to an older reference, the Human Genome build 19 (hg19) reference genome. Although there are now newer and slightly more accurate sequences, this sequence was a standard for references and is not terribly different from the more modern sequences used. Because it would not negatively affect our project, but would offer us more samples to pull from, we decided to switch our reference genome to hg19.

We obtained VCF files from the 1000 Genomes Project's final phase, phase 3. This phase contains 2504 samples representing 26 populations. From this, we could download a VCF file for each chromosome, containing information from each sample about SNPs. From these files, we could extract heterozygous SNPs.

To perform our analysis with global incidence and mortality rates, and to ensure we were pulling from female genomes, we needed demographic information from our samples, such as nationality and sex. These VCF files do not contain demographic information, so we sought out a public file that contained each sample title with their population (country), their superpopulation (continent), and their sex. Using this information, we were able to find the female samples from varying populations. We extracted these values to be used for analysis.

Analysis

To put the heterozygosity values into context, we decided to include controls. In every sequence we analyzed, we took two regions from the genome as controls. One sample was from chromosome 17, the location of BRCA1, and was the length of BRCA1. The other sample was from chromosome 13, the location of BRCA2, and was the length of BRCA2. We then found the number of heterozygous SNPs for these samples and divided the sum of the heterozygous SNPs from BRCA1 and BRCA2 by the sum of the heterozygous SNPs from the two controls.

Since the combined lengths of both experimental and control groups were equal, there was no normalization needed. This ratio was then used as our metric for heterozygosity in that sample. The purpose of these controls was to put the level of heterozygosity of the experimental genes into the context of the overall genome to see if heterozygosity was increased in the BRCA1 and BRCA2 genes.

To analyze this data, we created a series of scatterplots. We first analyzed breast cancer incidence rates. Using incidence rates from the IARC, we graphed the heterozygosity ratio from each sample on the x-axis and the incidence rate from the country that that sequence came from on the y-axis. We then calculated the average ratio per country and graphed the heterozygosity vs incidence using the country averages. Using mortality rates from the IARC, we followed this same architecture to produce two graphs on heterozygosity vs mortality. The scatterplots followed the color scheme of the IARC incidence and mortality graphs per continent, with Asia being represented as green, Africa represented as red, Latin America represented as yellow, and Europe represented as blue.

Analysis with HDI

HDI can have a significant impact on healthcare, including incidence and mortality of cancer. There are many factors that cause cancer, and we cannot include all of them in our analysis. However, we decided to include HDI as it was an accessible value that we could use for quantitative analysis. We extracted HDI values for the countries represented in our data and created 4 separate scatterplots with 4 levels of HDI.

These levels were split into very high (≥ 0.80), high (≥ 0.70), medium (≥ 0.55), and low (< 0.55). We created scatterplots for the countries that fit into their respective HDI categories to compare countries within the same realm of healthcare access. We did this to attempt to limit the possible effects of healthcare access on our data. We created 4 graphs of heterozygosity vs incidence at each level of HDI following the above process, as well as 4 graphs of heterozygosity vs mortality at each level of HDI following the above process.

Our final point of analysis was correlation coefficients. We calculated correlation coefficients for each data set in each scatterplot we created. We used these values to examine the statistical significance between heterozygosity and incidence/mortality on an individual and global scale.

RESULTS

Part 1

In the first part of the project, with 78 individuals sampled, we found a weak, positive correlation between heterozygosity in the BRCA1 gene and incidence rate. For Part 1 of the project, heterozygosity was measured as the number of heterozygous SNPs found in an individual's BRCA1 gene. The incidence rate was measured as the absolute number of breast cancer cases per 100,000 people per year, at an age-standardized rate. We will define two types of scatter plots for our project. Scatter plot Type 1 is where each individual is represented as its own data point. Scatter plot Type 2 is where the average of all individuals belonging to the same country is a data point. The correlation coefficient of Type 1 scatter plot is 0.1577. The correlation coefficient of Type 2 scatter plot is 0.2343 (Figure 1). The same analysis was performed for heterozygosity and mortality rate, with heterozygosity measured in the same manner. The mortality rate was measured as the absolute number of deaths per 100,000 people per year, at an age-standardized rate. There was a weak, negative correlation between heterozygosity in the BRCA1 gene and mortality rate. The correlation

coefficient of Type 1 scatter plot is -0.0249. The correlation coefficient of Type 2 scatter plot is -0.1116 (Figure 2). The scatter plots are color-coded by continent, consistent with the data on IARC. African countries are represented by red points; Asian countries are represented by green points; Latin American countries are represented by yellow points; European countries are represented by blue points.

These correlation coefficients were very weak. Due to the limited scope of our project at this point, with few samples, no control sequences, and one experimental gene, we could not accept these values as results for our data, as there were too many outlying factors. Finding low correlation is an acceptable result when there is enough data to support it, but we did not have enough data. This lead us to part 2 of the project, where we increased our scope by adding more samples, including the BRCA2 gene, and adding a control region.

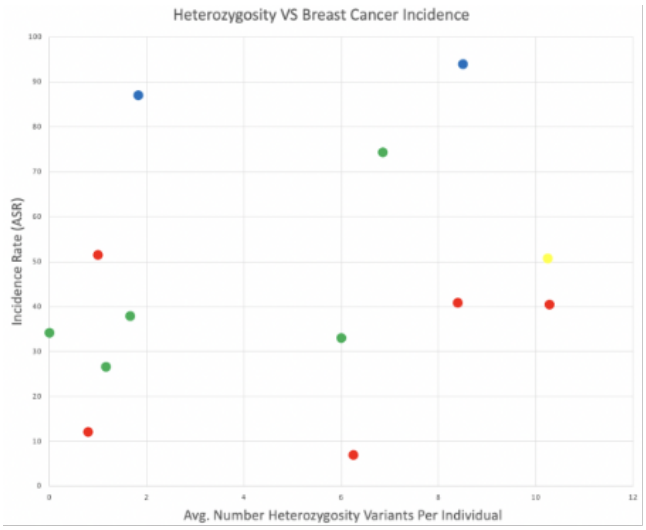

Fig. 1. Heterozygosity VS Incidence Type 2 scatter plot from Part 1.

Part 2

For the second part of the project, we sampled 1,187 individuals and evaluated both their BRCA1 and BRCA2 genes. For Part 2 of the project, heterozygosity was measured as an individual's heterozygosity ratio. Heterozygosity ratio was calculated as the number of heterozygous SNPs found in the individual's BRCA1 and BRCA2 genes over the number of heterozygous SNPs found in the two control regions in chromosome 17 and chromosome 13, with a combined length equal to that of the combined length of BRCA1 and BRCA2. This ratio was used because it put the heterozygosity of the cancer-related genes into the context of the rest of the genome, allowing for more accurate analysis.

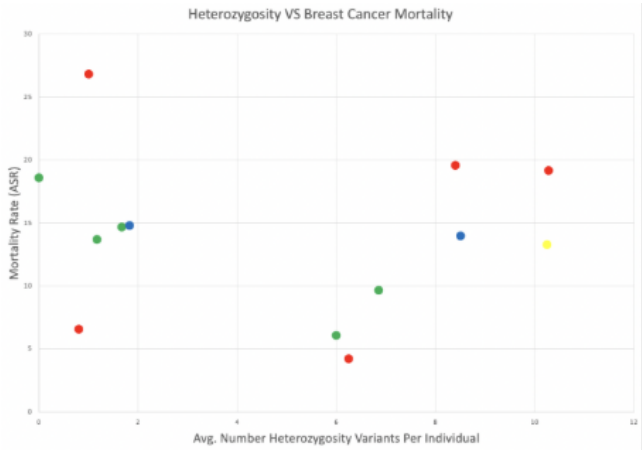

Fig. 2. Heterozygosity VS Mortality Type 2 scatter plot from Part 1.

Heterozygosity in the Control Region

The control region for chromosome 17 had an average of 53.9731 heterozygous SNPs across all individuals. The control region for chromosome 13 had an average of 146.0245 heterozygous SNPs. The BRCA1 gene region on chromosome 17 had an average of 71.9831 heterozygous SNPs. The BRCA2 gene region on chromosome 13 had an average of 95.2951 heterozygous SNPs.

Heterozygosity and Breast Cancer Rates

Both incidence rate and mortality rates were obtained from the IARC. We found a weak, negative correlation between heterozygosity and incidence rate in Part 2. The correlation coefficient of Type 1 scatter plot is -0.0501. The correlation coefficient of Type 2 scatter plot is -0.2506 (Figure 3). We found a weak, negative correlation between heterozygosity and mortality rate. The correlation of coefficient of Type 1 scatter plot is -0.0454. The correlation coefficient of Type 2 scatter plot is -0.2429 (Figure 4).

Though these correlation coefficients are still not very strong, they are backed by a more solid foundation of data, allowing us to feel confident that these values are significant to our research as opposed to a possible result of lack of data.

Correlation by HDI

The IARC holds HDI data for each country. We categorized these values into 4 groupings, very high, high, medium, and low. We calculated the correlation coefficients between heterozygosity vs incidence and heterozygosity vs mortality for each category of HDI on the individual level (Table 1). These values help us gain insight on the effects of access to healthcare on the incidence and mortality of breast cancer, an important aspect to cancer treatment.

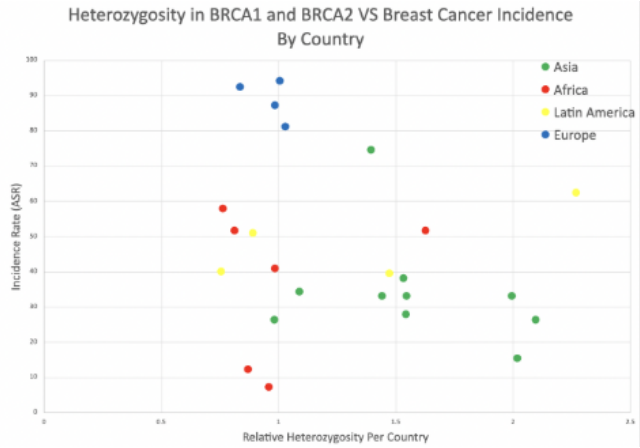

Fig. 3. Heterozygosity VS Incidence Type 2 scatter plot from Part 2.

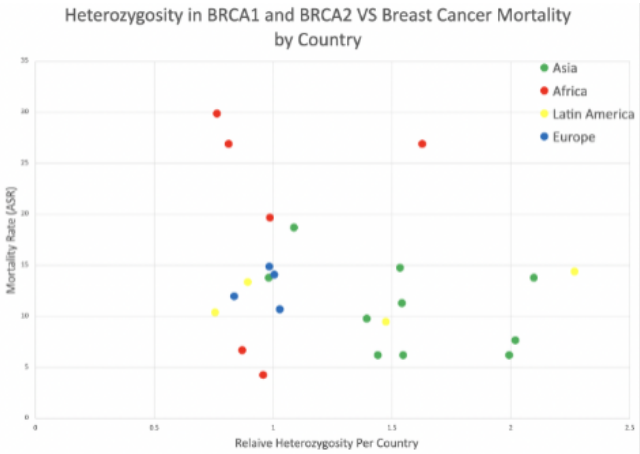

Fig. 4. Heterozygosity VS Mortality Type 2 scatter plot from Part 2.

	Low	Medium	High	Very-High
Het. vs Incidence	0.1068	-0.1583	-0.2022	-0.1334
Het. vs Mortality	0.107	-0.1587	-0.1658	0.0302

Fig. 5. Type 1 correlation coefficients by HDI category.

DISCUSSION

Part 1 vs Part 2

Across individuals on a global scale, there appears to be an overall weak, negative correlation between heterozygosity and incident rates and a weak, negative correlation between heterozygosity and mortality rates. We acknowledge in Part 1 there was a weak positive correlation between heterozygosity and incidence. However, we believe our results from Part 2 are more reliable. This is due to our increased number of individuals evaluated (1,187 in Part 2 and 78 in Part 1). Further, our measure of heterozygosity in Part 2 is more meaningful due to our use of control regions. The control regions were meant to represent a baseline level of heterozygosity expected in an individual's genome. Across all samples, there appears to be higher amounts of heterozygosity in chromosome 13 than in chromosome 17. There were more heterozygous SNPs in BRCA1 than there were in the control region of chromosome 17. In contrast, there were more heterozygous SNPs in the control region of chromosome 13 than there were in BRCA2. This may indicate that the levels of heterozygosity in the BRCA1 and BRCA2 genes are not significantly different from levels of heterozygosity in the rest of the genome, but more data is needed in order to make this claim.

Hypothesis Evaluation

Our hypothesis claimed that we would expect to see highest levels of heterozygosity rates in Africans, followed by Americans, South Asians, Europeans, and East Asians. The data we analyzed resulted in heterozygous ratios from highest to lowest order in a different order: East Asians (1.5794), South Asians (1.5301), Americans (1.4023), Africans (1.0183), and Europeans (0.9606). Furthermore, our hypothesis claimed that we would expect breast cancer rates to be in order from lowest to highest as: Africans, Americans, South Asians, Europeans, and East Asians. In our data, breast cancer incidence rates from lowest to highest are ordered as: South Asians (26.0711), Africans (36.8961), East Asians (41.6623), Americans (49.1818), and Europeans (88.5037). The European incidence rate is significantly higher than other geographic regions, potentially because of increased access to healthcare in these regions, with many European countries having a universal healthcare policy. We hypothesized that there would be a negative correlation between heterozygosity and breast cancer incidence and mortality rates

In our data, breast cancer mortality rates from lowest to highest are ordered as: East Asians (8.5176), Americans (12.0636), Europeans (12.7554), South Asians (13.0419), and Africans (19.0072). In terms of mortality rates more so than incidence rates, the order of lowest to highest mortality rate appears to correspond more to each region's HDI. Regions with overall higher country HDI's, such as East Asia and Europe, have lower rates of mortality compared to regions with overall lower country HDI's, such as South Asia and Africa, have higher rates of mortality. HDI is evaluated based on Standard of Living (Gross National Income per capita) and Health (life expectancy at birth), amongst other factors. An individual's income, the affordability of healthcare in a country, and the quality of breast cancer treatments in a healthcare system are all factors that could impact mortality rates in a population.

Analysis of Results

Our results show that for Low HDI countries, there is a weak, positive correlation between heterozygosity vs incidence and heterozygosity vs mortality rates. At higher levels of HDI (Medium, High, and Very-High), overall there is a weak, negative correlation between heterozygosity vs incidence and heterozygosity vs mortality rates. However, there is one exception: at the highest level of HDI there is a positive correlation coefficient close to zero between heterozygosity and mortality rates.

We acknowledge that there are many factors that impact breast cancer rates that are not evaluated in this research. However, we propose a few reasons as to why there are discrepancies from the overall negative correlation between heterozygosity and incidence and heterozygosity and mortality. Low HDI countries often have rural, isolated populations which can be characterized by limited genetic diversity and consequently lower heterozygosity rates. Urbanization can contribute to higher heterozygosity rates [11]. Therefore, in the urban populations of Low HDI countries, there may be higher rates of heterozygosity along with better access to healthcare for diagnosis and treatment found in cities. With greater discrepancies in healthcare access and income between rural and urban populations in Low HDI countries, this could be a contributing factor to the positive correlations seen here.

In terms of the positive correlation coefficient of 0.0302 between heterozygosity and mortality in Very-High HDI countries, this seems to indicate that the mortality rate amongst individuals of these countries has almost no relationship to the individual's genetic heterozygosity. While it's difficult to quantify how much, if any, impact genetic heterozygosity has on an individual's breast cancer survival, we believe that this 0.0302 correlation supports the claim that quality healthcare can meaningfully improve likelihood of survival. In countries with Very-High HDI it is more common practice to perform mammography screening, Magnetic Resonance Imaging (MRI) for high-risk groups, and genetic testing, all of which allow for earlier detection of breast cancer and even lead to preventative treatment.

Earlier detection and treatment can improve breast cancer survivability rates. Countries with high HDIs tend to have policies that allocate more resources for cancer research, treatment, and awareness. Therefore, our research suggests a negative correlation between heterozygosity and breast cancer rates and emphasizes the importance of early detection and treatment.

CONCLUSION

In this study, we investigated genetic heterozygosity and its relationship with breast cancer incidence and mortality across global populations. Across all individuals analyzed, our findings indicate a weak, negative correlation between heterozygosity and both incidence and mortality rates. Although Part 1 of our study revealed a weak, positive correlation between heterozygosity and incidence, we consider the results from Part 2 more robust due to the substantially larger sample size and the use of control regions to provide a more meaningful measurement of genome heterozygosity.

Our analysis of heterozygosity across BRCA1 and BRCA2 genes compared to control regions suggests that heterozygosity levels within these genes may not differ dramatically from the broader genome, although further data are needed to confirm this. We also observed higher heterozygosity on chromosome 13 than chromosome 17, from both the control and gene regions.

The relationship between heterozygosity and breast cancer incidence and mortality rates correlated to our hypothesis, but the order of expected heterozygosity differed from our hypothesis. For instance, Europe exhibited the highest incidence rates, likely reflecting increased access to healthcare and widespread cancer screening programs. Additionally, mortality rates more closely aligned with regional HDI levels than incidence did. Populations from higher HDI regions (e.g. Europe, East Asia) displayed lower mortality rates, while those in lower-HDI regions (e.g. South Asia, Africa) showed higher mortality. This is consistent with disparities between low and high HDI countries in terms of medical infrastructure, treatment quality, and early detection efforts.

Overall, our findings suggest a modest but consistent negative association between genetic heterozygosity and breast cancer incidence and mortality, on a global scale. However, the relationship is by no means deterministic, as our result also underscore the substantial influence of socioeconomic and healthcare variables complicate this relationship across global populations. These findings emphasize the crucial role of early detection and access to healthcare in reducing mortality. Further research to better understand breast cancer susceptibility from genetic differences across global regions may include analyzing a broader set of genetic markers across the genome and broadening the evaluated set of genes.

COMPETING INTERESTS

No competing interest is declared.

ACKNOWLEDGMENTS

The authors thank Michael Schatz and Mahler Revsine for their guidance and help with this project.

REFERENCES

1. Samuels, D.C., Wang, J., Ye, F., He, J., Levinson, R.T., Sheng, Q., Zhao, S., Capra, J.A., Shyr, Y., Zheng, W., and Guo, Y. Heterozygosity

Ratio, a Robust Global Genomic Measure of Autozygosity and Its Association with Height and Disease Risk. *Genetics*, 204(3):893–904, 2016. doi:10.1534/genetics.116.189936.

2. Xu, K., Kosoy, R., Shameer, K., et al. Genome-wide analysis indicates association between heterozygote advantage and healthy aging in humans. *BMC Genetics*, 20(52), 2019. doi:10.1186/s12863-019-0758-4.
3. Amos, W. Variation in Heterozygosity Predicts Variation in Human Substitution Rates between Populations, Individuals and Genomic Regions. *PLOS ONE*, 8(4):e63048, 2013. doi:10.1371/journal.pone.0063048.
4. National Center for Biotechnology Information. BRCA1 DNA repair associated [Homo sapiens]. National Institutes of Health, 2025. `https://www.ncbi.nlm.nih.gov/gene/672`.
5. National Cancer Institute. GDC Data Portal. National Institutes of Health, 2025. `https://portal.gdc.cancer.gov/analysis_page?app=CohortBuilder&tab=general`.
6. National Center for Biotechnology Information. Genome assembly GRCh38. National Institutes of Health, 2025. `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/`.
7. IGSR: The International Genome Sample Resource. Sample Data Portal, 2025. `https://www.internationalgenome.org/data-portal/sample`.
8. International Agency for Research on Cancer. Cancer Today: DataViz, 2025. `https://gco.iarc.fr/today/en/dataviz/strip-chart?mode=population&sexes=2&cancers=20`.
9. Stefansson, O.A., and Esteller, M. BRCA1 as a tumor suppressor linked to the regulation of epigenetic states: keeping oncomiRs under control. *Breast Cancer Research*, 14(2):304, 2012. doi:10.1186/bcr3119.
10. World Health Organization. Breast Cancer Fact Sheet, 2025. `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`.
11. Rudan, I., Carothers, A.D., Polasek O., Hayward, C., Vitart, V., Biloglav, Z., et al. Quantifying the increase in average human heterozygosity due to urbanisation. *European Journal of Human Genetics*, 16(9):1097–1102, 2008. doi:10.1038/ejhg.2008.48.