Repurposing Non-invasive prenatal testing data; population study of single nucleotide variants associated with colorectal cancer and Lynch syndrome

Natalia Forgacova (natali.forgacova@gmail.com)
Comenius University in Bratislava Faculty of Natural Sciences: Univerzita Komenskeho v Bratislave Prirodovedecka fakulta
https://orcid.org/0000-0001-6764-0895
Juraj Gazdarica
Comenius University Science Park, Bratislava
Jaroslav Budis
Comenius University Science Park, Bratislava
Jan Radvanszky
Comenius University Science Park, Bratislava
Tomas Szemes
Comenius University Science Park, Bratislava

Research Article

Keywords: Colorectal cancer, Lynch syndrome, Non-invasive prenatal testing, Low-coverage massively parallel whole genome sequencing

Posted Date: March 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-286309/v1

Abstract

PURPOSE: In our previous work, we described genomic data generated through non-invasive prenatal testing (NIPT) based on low-coverage massively parallel whole-genome sequencing of total plasma DNA of pregnant women in Slovakia as a valuable source of population specific data. In the present study, we used these data to determine the population allele frequency of common risk variants located in genes associated with colorectal cancer (CRC) and Lynch syndrome (LS).

METHODS: Allele frequencies of identified variants were compared with six world populations to detect significant differences between populations. Finally, we interpreted variants and searched for functional consequences and clinical significance of variants using publicly available databases.

RESULTS: Although we could not identify any pathogenic variants associated with CRC or LS in the Slovak population using NIPT data, we observed significant differences in the allelic frequency of risk CRC variants previously reported in GWAS and common variants located in genes associated with LS.

CONCLUSION: As Slovakia is the third country with the highest incidence of CRC per 100 000 population in the world, we highlight a need for studies dedicated to the cause of such a high incidence of CRC in Slovakia. We also assume that extensive cross-country data aggregation of NIPT results would represent an unprecedented source of information about human genome variation, also in cancer research.

Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide and the second most common cancer in Europe, causing an estimated 9,4 % of all cancer deaths in Europe. CRC is also a serious societal problem in Slovakia, with an incidence rate of 15,7 % and a mortality rate of 15,6 % (Bray et al. 2018; *Global Cancer Observatory* - GLOBOCAN 2020 data). Many risk factors and causes are associated with the likelihood of developing CRC, but the main reason is still not fully understood. The considerable geographical variability suggests that CRC is a complex polygenic disease caused by genetic and environmental factors and their interactions. Age, sex, lifestyle, and dietary habits (Rawla et al. 2019; Thanikachalam and Khan 2019), including meat and alcohol consumption (Cai et al. 2014; Dashti et al. 2017), tobacco smoking (Botteri et al. 2008; Limsui et al. 2010; Ordóñez-Mena et al. 2018), obesity and lack of physical activity (Gharahkhani et al. 2019; Rawla et al. 2019; Thrift et al. 2015) play a major role in the pathogenesis of CRC. Other well-known risk factors may also be inflammatory bowel diseases, acromegaly, renal transplantation with long-term immunosuppression, diabetes mellitus and insulin resistance, cholecystectomy, or androgen deprivation therapy (Rawla et al. 2019; Thanikachalam and Khan 2019).

Beside these, inherited susceptibility plays a significant role in the etiology of CRC because it can be responsible for about 35 % of all cases of colorectal cancer. However, high-penetrance germline variants in known genes (*APC, BRCA2, KRAS, NTS, SMAD4, POLE, BRAF, BMPR1A, POLD1, STK11, MUTYH* and DNA mismatch repair genes), which are associated with severe hereditary syndromes, such as familial adenomatous polyposis and Lynch syndrome (also called hereditary non-polyposis colorectal cancer), account for only 5-7 % of total CRC cases (Dekker et al. 2019; Rawla et al. 2019). Therefore, the remaining unknown heritability is probably explained by the interaction of common, low-penetrance variants identified through genome-wide association studies (GWAS). GWAS in ethnic/racial minority populations offers the opportunity to uncover genetic susceptibility factors and discover new genomic regions and loci that contribute risk for CRC development. Since 2007, more than 100 common risk variants have been successfully identified in GWAS, which have helped to elucidate the etiology of CRC (Al-Tassan et al. 2015; Lu et al. 2019; Schmit et al. 2019; Takahashi et al. 2017; Zeng et al. 2016; B. Zhang et al. 2014).

Lynch syndrome (LS) is an autosomal dominant hereditary cancer syndrome that accounts for approximately 3 % of all colorectal cancer cases (Biller et al. 2019). From a clinical point of view, 10-82 % (Yurgelun and Hampel 2018) of LS cases are associated with a lifetime risk of developing CRC, unless the risk is significantly lower in other types of cancer (Møller et al. 2017, 2018). LS is caused by pathogenic germline mutations in a class of genes called DNA mismatch repair (MMR) genes, mainly *MLH1*, located in 3p22.2 chromosome, and *MSH2*, located in 2p21 chromosome (Soares et al. 2018), which

represent 70-85% of cases of LS (Cox et al. 2018). Mutations found in *MSH6* (2p16.3), *PMS2* (7p22.1) (Le et al. 2017) and *MLH3* genes have lower incidence (Peltomäki 2016). Molecular investigations have also shown that *MSH3* (Duraturo et al. 2011) and germline 3´deletions of the *EPCAM* gene, which lead to epigenetic silencing of *MSH2* (Kuiper et al. 2011), are also implicated in the pathogenesis of LS. As a consequence of MMR pathway inactivation and loss of expression of MMR proteins, DNA replication errors accumulate typically resulting in microsatellite instability (MSI), which is generally detected in LS patients' tumor tissues (Shah et al. 2010). The diagnosis of LS involves three main steps, identification of patients and their familial history that meet the Amsterdam or Bethesda guidelines, presence of MSI in tumors and immunohistochemical analysis (IHC) of MMR protein expression. A definitive diagnosis of LS must be confirmed by detecting the germline mutations in MMR genes (Martin-Morales et al. 2018).

Non-invasive prenatal testing (NIPT) based on low-coverage massively parallel whole-genome sequencing of plasma DNA from pregnant women generates a large amount of data that provides the resources to investigate human genetic variations in the population. In our previous studies, we described the re-use of the data from NIPT for genome-scale population specific frequency determination of small DNA variants (Budis et al. 2019) and CNVs (Pös et al. 2019). Since pregnant women represent a relatively standard sample of the local female population, we assumed this NIPT data could also be used in the population study of CRC, the most common cancer in Slovakia. Some research concerning on genomic analysis of plasma from NIPT has also demonstrated NIPT data's efficiency and utility for viral genetic studies (Liu et al. 2018), genetic profiling of Viatnamese population (Tran et al. 2020) or detection of CNV aberrations (Pös et al. 2019; Pös et al. 2019).

The main aim of our study was a detailed analysis of common variants (MAF>0,05) that showed evidence of association with CRC in GWAS datasets and characterization of population variability from data generated by NIPT. We assumed that the genetic factors, mainly the increased specific population frequency of CRC and LS variants could be responsible for the high incidence of CRC in Slovakia. To test this hypothesis, allele frequencies of risk CRC variants identified in the Slovak population were compared with allele frequencies of risk CRC variants in 6 worldwide populations. As LS is among the most common hereditary CRC syndromes, the aim of our study was also to analyze population allele frequencies and describe clinical impacts of relevant variants located in known LS predisposing genes. To our knowledge, this was the first population study of CRC using NIPT data conducted exclusively in the Slovak population.

Methods

Data source

The laboratory procedure used, to generate the NIPT data, were as follows: DNA from plasma of peripheral maternal blood was isolated for NIPT analysis from 1,501 pregnant women after obtaining a written informed consent consistent with the Helsinki declaration from the subjects. The population cohort consisted from women in reproductive age between 17-48 years with a median of 35 years. Genomic information from a sample consisted of maternal and fetal DNA fragments. Each included individual agreed to use their genomic data in an anonymized form for general biomedical research. The NIPT study (study ID 35900_2015) was approved by the Ethical Committee of the Bratislava Self-Governing Region (Sabinovska ul.16, 820 05 Bratislava) on 30th April of 2015 under the decision ID 03899_2015. Blood samples were collected to EDTA tubes and plasma was separated in dual centrifugation procedure. DNA was isolated from 700 µl of plasma using DNA Blood Mini kit (Qiagen, Hilden, DE) according to standard protocol. Sequencing libraries were prepared from each sample using TruSeq Nano kit HT (Illumina, San Diego, CA, USA) following standard protocol with omission of DNA fragmentation step. Individual barcode labelled libraries were pooled and sequenced using low-coverage whole-genome sequencing on an Illumina NextSeq500 platform (Illumina, San Diego, CA, USA) by performing paired end sequencing of 2×35 bases (Minarik et al. 2015).

Data analysis

Called variants, that support the findings of this study, are available in VCF format from https://sites.google.com/view/snipt under a flag “SNIPT”. Identified variants with corresponding information were submitted to dbSNP under the following identifiers; Handle: BIOINF_KMB_FNS_UNIBA, Batch id: 1062867 (Budis et al., 2019).

Analyses of common variants previously reported to be risk variants for CRC

We combined genotype data from all previously reported GWAS studies available online (https://www.gwascentral.org/) for the years 2007-2020, specifically 66 GWAS studies of CRC risk variants that included individuals with European, Asian and African American ancestry. Using data from these GWAS datasets, we identified 116 risk variants associated with CRC, which were then merged with our data of identified variants from NIPT. Risk variants that were not found in NIPT data were excluded from the analysis. All identified variants in the Slovak population used for further analyses were common (MAFs > 0.05). Subsequently, allele frequencies of CRC risk variants for each population (East Asian, South Asian, African, American, Finnish European and non-Finnish European) were extracted from the gnomAD database available online (v3.0, downloaded from https://gnomad.broadinstitute.org/downloads) and compared with our frequencies determined for the Slovak population from NIPT data. Allele frequency in each population and allele frequency differences were plotted using boxplots and PCA analysis using matplotlib.pyplot library. Outliers of boxplots that represent variants with highly different frequencies between Slovak and non-Finnish populations were annotated via published literature and studies (in dbSNP (https://www.ncbi.nlm.nih.gov/snp/) and GWAS (https://www.gwascentral.org/)).

Analyses of variants located in genes associated with LS

After analyzing variants associated with CRC, we focused on the study of variants associated with LS. First, we filtered out a group of variants located in 7 genes known to be associated with LS *(MLH1, PMS2, MSH6, MLH3, MSH2, TGFBR2, EPCAM*). The genomic locations of genes were determined by the GeneCards database (https://www.genecards.org/). From the dataset of identified variants in LS associated genes, we excluded variants in low complexity genomic regions (soft masked in the reference FASTA file), eliminating the variants that could represent sequencing artifacts or repetitive regions. All variants that were used for further analysis were annotated using Ensembl Variant Effect Predictor (VEP, version 101_GRCh38). In our dataset, based on ClinVar database annotation of the most common types of pathogenic and likely pathogenic variants associated with LS (https://www.ncbi.nlm.nih.gov/clinvar), we selected variants including frameshift, missense, nonsense, splice site, non-coding and UTR variants. After this filtering, allele frequencies for both groups of variants (all variants identified in LS genes and selected types of variants) for each population (East Asian, South Asian, African, American, Finnish European and non-Finnish European) were extracted from the gnomAD database (v3.0, downloaded from https://gnomad.broadinstitute.org/downloads) and compared with our frequencies determined for the Slovak population from NIPT data. Allele frequency in each population and allele frequency differences were plotted using boxplots and PCA analysis using matplotlib.pyplot library. Outliers of boxplots representing variants with allele frequency differences more than 10% were annotated via published literature and studies (in dbSNP (https://www.ncbi.nlm.nih.gov/snp/) and GWAS (https://www.gwascentral.org/)).

Results

Analyses of common variants previously reported to be risk variants for CRC

In the analysis of the 66 GWAS studies that included all identified risk variants associated with colorectal carcinogenesis from 2007-2020, we investigated 116 variants. There were 25 independent CRC risk variants in Asian population, 62 risk variants found in European population, 27 risk variants in both European and Asian population and 2 risk variants located in African American population that were previously reported in GWAS (Table 1).

After merging all identified variants from GWAS (116 risk variants) with our NIPT data, we identified 106 common risk CRC variants (Supplementary table 1), while 10 risk variants that were not called in the Slovak population were excluded from further analysis. The allele frequencies of 106 variants identified in our population sample (Slovak population) and the allele

frequencies of variants for 6 world populations (East Asian, South Asian, African, American, Finnish European and non-Finnish European) obtained by gnomAD database (Supplementary table 2) are shown in graphical comparison by Boxplots (Figure 1) and principal component analysis (PCA) (Figure 2). As shown in Figure 1, the MAF ranged from 0.0-0.963109 in 6 world populations that is comparable with the Slovak population (0.0521-0.931). The median allele frequency for the Slovak population reached the value of 0.4072, which is closest to the value of the median of the American population (MED=0.3991) and Finnish population (MED=0.4166). PCA placed our sample set most closely to the two gnomAD population sample sets, i.e., to the Finnish and non-Finnish European population.

Next, we compared known allele frequencies of 106 CRC risk variants in our sample set from the Slovak population to allele frequencies of CRC variants in six world populations. The final findings of allele frequency differences are shown in Figure 3. The median allele frequency for comparing the Slovak population and non-Finnish European population reached the value of 0.002285. Together, we identified 14 variants, whose difference in population allele frequency was more than 10%. Table 2 includes annotation information about these variants by dbSNP NCBI, ClinVar database and population comparison in which they were identified.

Analyses of variants located in genes associated with LS

In the analysis of LS, we identified 1212 variants in our sample set from NIPT that were located in genes known to be associated with LS, i.e., *MLH1*, *PMS2*, *MSH6*, *TGFBR2*, *MLH3*, *MSH2* and *EPCAM*. After excluding variants from low complexity regions, we obtained 648 variants that were finally annotated by VEP and used for further analysis. The allele frequencies of 648 variants identified in our population sample (Slovak population) and the allele frequencies of variants for 6 world populations (East Asian, South Asian, African, American, Finnish European and non-Finnish European) obtained by gnomAD database (Supplementary table 3) are shown in graphical comparison by Boxplots (Figure 4) and principal component analysis (PCA) (Figure 5). As shown in Figure 4, the MAF ranged from 0.0-1.0 in 6 world populations. In the Slovak population, all variants were with MAF > 0.05 (0.0502-1.0). The median allele frequency for the Slovak population reached the value of 0.2204, which is closest to the value of the median of the South Asian population (MED=0.221274). PCA placed our sample set most closely to the non-Finnish European population (Figure 5).

In the next step, to identify variants having significantly different frequencies, we compared known allele frequencies of 648 variants located in genes associated with LS identified in our sample set from the Slovak population to allele frequencies of these variants in six gnomAD world populations. The final findings of allele frequency differences are shown in Figure 6. The median allele frequency for the comparison of the Slovak population and non-Finnish European population reached the value of -0.01093. By comparing the allele frequency of variants of the Slovak and non-Finnish populations, we identified a total of 64 outliers. Most outliers were found in the *MSH2* gene, others in *MSH6, TGFBR2, PMS2, MLH1* and *EPCAM*. We did not identify any outlying variant in the *MLH3* gene. The variation type of all outliers was “intronic variant” and the clinical significance of all outliers was not reported in ClinVar. All this annotation information, also including chromosome position, variants ID, reference and alternative allele, genes, is available in Supplementary table 4.

Our analysis also included allele frequency comparison of selected variants from 648 variants identified in our sample set of NIPT data. We focused on frameshift, missense, nonsense, splice site, non-coding and UTR variants, annotated in the ClinVar database as the most common types of pathogenic and likely pathogenic variants associated with LS. However, from these selected types of variants, we found only UTR and non-coding variants in our dataset of 648 variants. Other types of variants (downstream, upstream, and intron) were excluded from further analysis. Finally, we selected 18 variants, 10 UTR and 8 non-coding variants (all selected variants with annotation information by VEP and ClinVar are available in Table 3). We compared known allele frequencies of these 18 selected variants identified in our population sample (Slovak population) to the six gnomAD world populations. The final findings of allele frequency differences are shown in Figure 7. The median allele frequency for the comparison of the Slovak population and non-Finnish European population reached the value of 0.014215, which is closest to the value of the median allele frequency comparison of the South Asian and Slovak population (MED=0.014186). By comparing the allele frequency of variants of the Slovak and six gnomAD world population, we

identified a total of 4 outliers - rs10951973, rs10951972 (identified in Slovak-American population comparison), rs6791557 (in Slovak-American and Slovak-non-Finnish European population comparison) and rs9852378 (in Slovak-South Asian population comparison). All outliers were non-coding variants, rs10951973 and rs10951972 located in the *PMS2* and rs6791557 located in the *TGFBR2* were not reported in ClinVar. The rs9852378 SNP, detected in the *MLH1*, was reported as benign by ClinVar.

Finally, we analyzed allele frequencies of pathogenic and likely pathogenic variants associated with LS annotated in the ClinVar database. From 229 SNPs with pathogenic and likely pathogenic clinical significance, only 15 have non-zero AF records in the gnomAD database. As shown in Figure 8, all found AF are significantly below 5% (Table 4, Figure 8).

Discussion

Population genetic studies currently have a huge impact on the study of genomics (Beyene and Pare 2014). The detection of risk variants in a population and identifying their genetic relationships have advanced our understanding of the human genome's variability and led to the elucidation of many factors that influence cancer risk. In recent years, NGS technologies have played a key role in colorectal cancer research and have become a useful tool for cancer diagnostics and screening (Budis et al. 2019; Valle et al. 2019; Yurgelun et al. 2015; Zhu et al. 2018). Due to the high incidence of colorectal cancer in the Slovak population, it is crucial to determine the possible causes of the high incidence of this disease in Slovakia.

Non-invasive prenatal testing of common fetal chromosomal aberrations, using low-coverage massively parallel whole-genome sequencing of maternal plasma cell-free DNA (cfDNA) of pregnant women, has become the fastest low-cost genomic DNA test that is rapidly implemented in clinical practice. Currently, more than 3 million NIPT tests are carried out worldwide each year, and the large amount of data generated during NIPT provides the resources to investigate human genetic variations in the population (Budis et al. 2019). In our study, we analyzed low-coverage massively parallel whole-genome sequencing data of total plasma DNA from pregnant women generated for NIPT screening to characterize the variants in genes associated with CRC and LS in the Slovak population. To our knowledge, the present study is the first population analysis of CRC and LS variants worldwide and also in the Slovak population using NIPT data. We illustrate the utility of these genomic data for clinical genetics and population studies.

Over the past two decades, GWAS offer the opportunity to uncover genetic susceptibility factors for CRC and provide insights into the biological basis of CRC etiology. These studies have demonstrated that only a fraction of colorectal cancer heritability is explained by known risk-conferring genetic variation, whereas the remaining genetic risk of CRC may be accounted for by a combination of high-prevalence and low-penetrance of common genetic variants. To date, a large number of common genetic variants have been identified by the GWAS approach, which has intimately connected to the onset of CRC (Hofer et al. 2017; Jiao et al. 2014; Law et al. 2019; Lu et al. 2019; Schmit et al. 2019; Wang et al. 2017; K. Zhang et al. 2014).

By pooling GWAS data of risk variants associated with colorectal carcinogenesis from 2007-2020 and data variants in our population sample from NIPT, we have identified 106 common risk CRC variants. When we compared allele frequencies of these variants to allele frequencies in six gnomAD world population, finally 13 common risk variants were found that showed statistically significant differences in population allele frequencies - rs5934683, rs7252505, rs4779584, rs1535, rs174550, rs4246215, rs11196172, rs10904849, rs6928864, rs3131043, rs1476570, rs12659017, rs397775554.

The SNP rs5934683 is located on chromosome Xp22.2 between two genes, *GPR143* (G protein-coupled receptor 143), which is expressed by melanocytes and retinal pigment epithelium and *SHROOM2* (shroom family member 2), a human homolog of the *Xenopus laevis APX* gene that has important functions in cell morphogenesis including endothelial and epithelial tissue development (K. Zhang et al. 2014). Missense mutations in this gene have been detected in large-scale screens for recurring mutations in cancer cell lines. Both *GPR143* and *SHROOM2* play a role in melanosome biogenesis and retinal pigmentation. It is known that abnormal retinal pigmentation, similar to the congenital hypertrophy of retinal pigment

epithelium lesions, are typical of the familial adenomatous polyposis syndrome (FAP), one of the inherited syndromes of CRC (Closa et al. 2014). The relationship between Xp22.2 and CRC risk represents the first evidence for the role of X-chromosome variation in predisposition to non-sex-specific cancer (Dunlop et al. 2012).

The SNP rs7252505, located in the 19q13 risk locus, is in an intron of the gene *GPATCH1* (G-patch domain containing 1). Although *GPATCH1* is expressed in the colon, little is known about its function other than the fact that it contains a G-patch domain, a domain typically associated with RNA processing. One study found that rs7252505 was associated with CRC in African Americans (Wang et al. 2013, 2017).

Intergenic variant rs4779584 in chromosomal region 15q13.3 lies between *SCG5* and *GREM1,* and the association between this SNP to CRC has been identified in several GWAS studies (Hong et al. 2015; Lu et al. 2019).

The rs4246215 polymorphism is located in the FEN1 in the long arm of chromosome 11 (11q12.2). The association between this SNP and the potential risk of different types of cancers, including esophageal, lung, gastrointestinal, gallbladder, breast cancer in Chinese and Iran populations, glioma and childhood leukemia, has been previously studied. The rs4246215 variant was also associated with colorectal cancer in East Asians and the Chinese population (Chou et al. 2017; Moazeni-Roodi et al. 2019).

To identify variants that may predispose to LS and may cause the high incidence of CRC in Slovakia, we used NIPT data, including variants with at least 5% AF and coverage at least 100 reads per variant. To verify the reliability of the found variants using NIPT, we selected 15 variants with AF below 5% and validated them using Sanger sequencing. For this reason, it is not possible to find rare variants with AF under 5%. Initially, we selected gene variants known to be associated with LS and we focused on their population AF in gnomAD database and as well as on pathogenicity as reported in public database ClinVar. No publications are available for all variants showing statistically significant differences in population allele frequencies and selected 18 variants. The rs9852378 SNP was reported as benign by ClinVar, and other variants were not reported in ClinVar.

Our study has several key shortcomings. None of the variants identified in this study are pathogenic or likely pathogenic due to their extremely low frequency in the general population (see RESULTS, Figure 8). From the total number of pathogenic or likely pathogenic variants annotated in the ClinVar database, we could determine the population frequency of only 15 variants even when using the gnomAD database (see RESULTS, Table 4). Second, the sample size was relatively small and it is strongly biased towards females. We assume that even larger sample sets will also not offer opportunities to detect such low frequencies of LS variants in the population using NIPT data. Third, a substantial portion of identified variants was removed from analyses due to technical limitations, mainly because of their location in low complexity regions. Although these could be technical artifacts (Kubiritova et al. 2019), they could also be real variants having biological effects that are yet generally hardly determinable, but likely existing (Trost et al. 2020). Moreover, colorectal cancer is a disease caused by a combination of multiple genes and environmental factors. To assess the relationship between the variants identified in population and CRC development, it is very important in future research to study the interaction between genes and also the environment on the colorectal cancer risk. Although suitable for the determination of general population frequencies of independent variants, NIPT data are unsuitable for calculations (such as polygenic risk score determinations) based on exact combinations of these variants in individuals, which may *de facto* determine the risk of individuals to develop certain diseases.

The underlying mechanism for a high incidence of CRC in the Slovak population is still unclear at the moment; however, it is possible that genetic factors, like the most common inherited syndrome – LS, play a crucial role in colorectal etiology. We have performed a literature search in PubMed focused on population studies of CRC and LS in Slovakia from 2010-2020 using next-generation sequencing. In the Slovak population, only a few population studies of risk variants have been conducted to elucidate the etiology of CRC (Kašubová et al. 2019; Mahmood et al. 2014; Skerenova et al. 2017). In general, little is known about risk variants associated with CRC or LS in the Slovak population.

Identifying mutations associated with CRC in populations with high mortality rate, such as the Slovak population, is important to reduce the incidence of this multifactor disorder. The findings from these studies suggest a lack of understanding of the mechanism of many risk variants of CRC. Due to study limitations, we could not identify any pathogenic variants associated with LS in the Slovak population using NIPT data. On the other hand, NIPT data is not a major obstacle to better results, as pathogenic variants have extremely low frequencies in the general population. Even in most cases, the frequencies are not known. However, we identified several promising common risk variants associated with CRC previously reported in GWAS studies that represent variants with highly different frequencies between Slovak and non-Finnish populations in boxplots. Since NIPT expands rapidly to millions of individuals each year, the reuse of these data reduces the cost of large-scale population studies and likely provides an acceptable background for information about genomic variation. Finally, future population studies on larger sample sets with various types of mutations are needed to reveal new mechanisms of pathogenicity and links to new biological pathways, which may be useful in designing preventive strategies and treatment of CRC.

Declarations

FUNDING

This work was supported by the PANGAIA project H2020-MSCA-RISE-2019 [Grant agreement ID: 872539] funded under H2020-EU.1.3.3. Programme; by the OP Integrated Infrastructure for the project: Long term strategic research and development focused on the occurrence of Lynch syndrome in the Slovak population and possibilities of prevention of tumors associated with this syndrome [ITMS: 313011V578], co-financed by the European Regional Development Fund (ERDF); by the Operational Program Integrated Infrastructure [ITMS code 313011F988], co-financed by the ERDF and by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences (VEGA 1/0305/19).

Conflict of interest All authors declare that there is no confict of interest and they have seen and approved the manuscript submitted.

References

Al-Tassan NA, Whiffin N, Hosking FJ et al (2015) A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Sci Rep 5:10442. doi:10.1038/srep10442

Beyene J, Pare G (2014) Statistical genetics with application to population-based study design: a primer for clinicians. Eur Heart J 35:495–500. doi:10.1093/eurheartj/eht272

Biller LH, Syngal S, Yurgelun MB (2019) Recent advances in Lynch syndrome. Fam Cancer 18:211–219. doi:10.1007/s10689-018-00117-1

Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P (2008) Smoking and colorectal cancer: a meta-analysis. JAMA 300:2765–2778. doi:10.1001/jama.2008.839

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394–424. doi:10.3322/caac.21492

Budis J, Gazdarica J, Radvanszky J et al (2019) Non-invasive prenatal testing as a valuable source of population specific allelic frequencies. J Biotechnol 299:72–78. doi:10.1016/j.jbiotec.2019.04.026

Budis J, Kucharik M, Duris F et al (2019) Dante: genotyping of known complex and expanded short tandem repeats. Bioinformatics 35:1310–1317. doi:10.1093/bioinformatics/bty791

Cai S, Li Y, Ding Y, Chen K, Jin M (2014) Alcohol drinking and the risk of colorectal cancer death: a meta-analysis. Eur J Cancer Prev 23:532–539. doi:10.1097/CEJ.0000000000000076

Chou A-K, Shen M-Y, Chen F-Y et al (2017) The Association of Flap Endonuclease 1 Genotypes with the Susceptibility of Endometriosis. Cancer Genomics Proteomics 14:455–460. doi:10.21873/cgp.20055

Closa A, Cordero D, Sanz-Pamplona R et al (2014) Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. Carcinogenesis 35:2039–2046. doi:10.1093/carcin/bgu092

Cox VL, Saeed Bamashmos AA, Foo WC et al (2018) Lynch Syndrome: Genomics Update and Imaging Review. Radiographics 38:483–499. doi:10.1148/rg.2018170075

Dashti SG, Buchanan DD, Jayasekara H et al (2017) Alcohol Consumption and the Risk of Colorectal Cancer for Mismatch Repair Gene Mutation Carriers. Cancer Epidemiol Biomarkers Prev 26:366–375. doi:10.1158/1055-9965.EPI-16-0496

Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB (2019) Colorectal cancer. Lancet 394:1467–1480. doi:10.1016/S0140-6736(19)32319-0

Dunlop MG, Dobbins SE, Farrington SM et al (2012) Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat Genet 44:770–776. doi:10.1038/ng.2293

Duraturo F, Liccardo R, Cavallo A, De Rosa M, Grosso M, Izzo P (2011) Association of low-risk MSH3 and MSH2 variant alleles with Lynch syndrome: probability of synergistic effects. Int J Cancer 129:1643–1650. doi:10.1002/ijc.25824

Gharahkhani P, Ong J-S, An J et al (2019) Effect of increased body mass index on risk of diagnosis or death from cancer. Br J Cancer 120:565–570. doi:10.1038/s41416-019-0386-9

Global Cancer Observatory. [cited 9 Nov 2020]. Available: https://gco.iarc.fr/

Hofer P, Hagmann M, Brezina S et al (2017) Bayesian and frequentist analysis of an Austrian genome-wide association study of colorectal cancer and advanced adenomas. Oncotarget 8:98623–98634. doi:10.18632/oncotarget.21697

Hong SN, Park C, Kim J-I et al (2015) Colorectal cancer-susceptibility single-nucleotide polymorphisms in Korean population. Journal of Gastroenterology and Hepatology: 849–857. doi:10.1111/jgh.12331

Jiao S, Peters U, Berndt S et al (2014) Estimating the heritability of colorectal cancer. Hum Mol Genet 23:3898–3905. doi:10.1093/hmg/ddu087

Kašubová I, Kalman M, Jašek K et al (2019) Stratification of patients with colorectal cancer without the recorded family history. Oncol Lett 17:3649–3656. doi:10.3892/ol.2019.10018

Kubiritova Z, Gyuraszova M, Nagyova E et al (2019) On the critical evaluation and confirmation of germline sequence variants identified using massively parallel sequencing. J Biotechnol 298:64–75. doi:10.1016/j.jbiotec.2019.04.013

Kuiper RP, Vissers LELM, Venkatachalam R et al (2011) Recurrence and variability of germline EPCAM deletions in Lynch syndrome. Hum Mutat 32:407–414. doi:10.1002/humu.21446

Law PJ, Timofeeva M, Fernandez-Rozadilla C et al (2019) Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat Commun 10:2154. doi:10.1038/s41467-019-09775-w

Le S, Ansari U, Mumtaz A et al (2017) Lynch Syndrome and Muir-Torre Syndrome: An update and review on the genetics, epidemiology, and management of two related disorders. Dermatol Online J 23. Available: https://www.ncbi.nlm.nih.gov/pubmed/29447627

Limsui D, Vierkant RA, Tillmans LS et al (2010) Cigarette smoking and colorectal cancer risk by molecularly defined subtypes. J Natl Cancer Inst 102:1012–1022. doi:10.1093/jnci/djq201

Liu S, Huang S, Chen F et al (2018) Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. Cell 175:347–359.e14. doi:10.1016/j.cell.2018.08.016

Skerenova M, Halasova E, Matakova T et al (2017) Low Variability and Stable Frequency of Common Haplotypes of the TP53 Gene Region in Colorectal Cancer Patients in a Slovak Population. Anticancer Research 1901–1907. doi:10.21873/anticanres.11528

Lu Y, Kweon S-S, Tanikawa C et al (2019) Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer. Gastroenterology 156:1455–1466. doi:10.1053/j.gastro.2018.11.066

Mahmood S, Sivoňová M, Matáková T et al (2014) Association of EGF and p53 gene polymorphisms and colorectal cancer risk in the Slovak population. Open Medicine 405–416. doi:10.2478/s11536-013-0300-4

Martin-Morales L, Rofes P, Diaz-Rubio E et al (2018) Novel genetic mutations detected by multigene panel are associated with hereditary colorectal cancer predisposition. PLoS One 13:e0203885. doi:10.1371/journal.pone.0203885

Minarik G, Repiska G, Hyblova M et al (2015) Utilization of Benchtop Next Generation Sequencing Platforms Ion Torrent PGM and MiSeq in Noninvasive Prenatal Testing for Chromosome 21 Trisomy and Testing of Impact of In Silico and Physical Size Selection on Its Analytical Performance. PLoS One 10:e0144811. doi:10.1371/journal.pone.0144811

Moazeni-Roodi A, Ghavami S, Ansari H, Hashemi M (2019) Association between the flap endonuclease 1 gene polymorphisms and cancer susceptibility: An updated meta-analysis. Journal of Cellular Biochemistry 13583–13597. doi:10.1002/jcb.28633

Møller P, Seppälä T, Bernstein I et al (2017) Incidence of and survival after subsequent cancers in carriers of pathogenic MMR variants with previous cancer: a report from the prospective Lynch syndrome database. Gut 66:1657–1664. doi:10.1136/gutjnl-2016-311403

Møller P, Seppälä TT, Bernstein I et al (2018) Cancer risk and survival in carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database. Gut 67:1306–1316. doi:10.1136/gutjnl-2017-314057

Ordóñez-Mena JM, Walter V, Schöttker B et al (2018) Impact of prediagnostic smoking and smoking cessation on colorectal cancer prognosis: a meta-analysis of individual patient data from cohorts within the CHANCES consortium. Ann Oncol 29:472–483. doi:10.1093/annonc/mdx761

Peltomäki P (2016) Update on Lynch syndrome genomics. Fam Cancer 15:385–393. doi:10.1007/s10689-016-9882-8

Pös O, Budis J, Kubiritova Z et al (2019) Identification of Structural Variation from NGS-Based Non-Invasive Prenatal Testing. Int J Mol Sci 20. doi:10.3390/ijms20184403

Pös O, Budiš J, Szemes T (2019) Recent trends in prenatal genetic screening and testing. F1000Res 8. doi:10.12688/f1000research.16837.1

Rawla P, Sunkara T, Barsouk A (2019) Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Prz Gastroenterol 14:89–103. doi:10.5114/pg.2018.81072

Schmit SL, Edlund CK, Schumacher FR et al (2019) Novel Common Genetic Susceptibility Loci for Colorectal Cancer. J Natl Cancer Inst 111:146–157. doi:10.1093/jnci/djy099

Shah SN, Hile SE, Eckert KA (2010) Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res 70:431–435. doi:10.1158/0008-5472.CAN-09-3049

Soares BL, Brant AC, Gomes R et al (2018) Screening for germline mutations in mismatch repair genes in patients with Lynch syndrome by next generation sequencing. Fam Cancer 17:387–394. doi:10.1007/s10689-017-0043-5

Takahashi Y, Sugimachi K, Yamamoto K et al (2017) Japanese genome-wide association study identifies a significant colorectal cancer susceptibility locus at chromosome 10p14. Cancer Sci 108:2239–2247. doi:10.1111/cas.13391

Thanikachalam K, Khan G (2019) Colorectal Cancer and Nutrition. Nutrients 11. doi:10.3390/nu11010164

Thrift AP, Gong J, Peters U et al (2015) Mendelian Randomization Study of Body Mass Index and Colorectal Cancer Risk. Cancer Epidemiol Biomarkers Prev 24:1024–1031. doi:10.1158/1055-9965.EPI-14-1309

Tran NH, Vo TB, Nguyen VT et al (2020) Genetic profiling of Vietnamese population from large-scale genomic analysis of non-invasive prenatal testing data. Sci Rep 10:19142. doi:10.1038/s41598-020-76245-5

Trost B, Engchuan W, Nguyen CM et al (2020) Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature 586:80–86. doi:10.1038/s41586-020-2579-z

Valle L, de Voer RM, Goldberg Y et al (2019) Update on genetic predisposition to colorectal cancer and polyposis. Mol Aspects Med 69:10–26. doi:10.1016/j.mam.2019.03.001

Wang H, Haiman CA, Burnett T et al (2013) Fine-mapping of genome-wide association study-identified risk loci for colorectal cancer in African Americans. Hum Mol Genet 22:5048–5055. doi:10.1093/hmg/ddt337

Wang H, Schmit SL, Haiman CA et al (2017) Novel colon cancer susceptibility variants identified from a genome-wide association study in African Americans. Int J Cancer 140:2728–2733. doi:10.1002/ijc.30687

Yurgelun MB, Allen B, Kaldate RR et al (2015) Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome. Gastroenterology 149:604–13.e20. doi:10.1053/j.gastro.2015.05.006

Yurgelun MB, Hampel H (2018) Recent Advances in Lynch Syndrome: Diagnosis, Treatment, and Cancer Prevention. American Society of Clinical Oncology Educational Book 101–109. doi:10.1200/edbk_208341

Zeng C, Matsuda K, Jia W-H et al (2016) Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. Gastroenterology 150:1633–1645. doi:10.1053/j.gastro.2016.02.076

Zhang B, Jia W-H, Matsuda K et al (2014) Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nat Genet 46:533–542. doi:10.1038/ng.2985

Zhang K, Civan J, Mukherjee S, Patel F, Yang H (2014) Genetic variations in colorectal cancer risk and clinical outcome. World J Gastroenterol 20:4167–4177. doi:10.3748/wjg.v20.i15.4167

Zhu L, Huang Y, Fang X et al (2018) A Novel and Reliable Method to Detect Microsatellite Instability in Colorectal Cancer by Next-Generation Sequencing. J Mol Diagn 20:225–231. doi:10.1016/j.jmoldx.2017.11.007

Tables

Table 1. 116 risk variants associated with CRC identified in 66 GWAS studies from 2007-2020.

rs_ID	CHR*	POS*	POP*	GENE	REFERENCE	PMID
rs7252505	19q13	33,084,158	AFR A	GPATCH1	Wang et al. 2017	28295283
rs56848936	19q13.3	46,321,507	AFR A	SYMPK	Wang et al. 2017	28295283
rs7542665	1p31.3	62,673,037	ASN	L1TD1	Lu et al. 2019	30529582
rs12143541	1p32.3	55,247,852	ASN	TTC22	Law et al. 2019	31089142
rs201395236	1q44	245,181,421	ASN	EFCAB2	Lu et al. 2019	30529582
rs7606562	2p16.3	48,686,695	ASN	PPP1R21	Lu et al. 2019	30529582
rs113569514	3q22.2	133,748,789	ASN	SLCO2A1	Lu et al. 2019	30529582
rs12659017	5q23.2	125,988,175	ASN	ALDH7A1, PHAX	Lu et al. 2019	30529582
rs639933	5q31.1	134,467,751	ASN	C5orf66, LOC105379188	Law et al. 2019	31089142
rs647161	5q31.1	134,499,092	ASN	PITX1	Jia et al. 2013	23263487
rs6933790	6p21.1	41,672,769	ASN	TFEB	Law et al. 2019	31089142
rs4711689	6p21.1	41,692,812	ASN	TFEB	Zeng et al. 2016	26965516
rs6906359	6p21.31	35,528,378	ASN	FKBP5	Schmit et al. 2018	29917119
rs3830041	6p21.32	32,191,339	ASN	NOTCH4	Lu et al. 2019	30529582
rs6584283	10q24.2	101,290,301	ASN	NKX2–3	Lu et al. 2019	30529582
rs77969132	12p11.21	31,594,813	ASN	DENND5B	Lu et al. 2019	30529582
rs11064437	12p13.31	6,982,162	ASN	SPSB2	Zeng et al. 2016	26965516
rs2730985	12q12	43,130,624	ASN	PRICKLE1	Lu et al. 2019	30529582
rs1886450	13q22.1	73,986,628	ASN	KLF5, KLF12	Lu et al. 2019	30529582
rs4341754	16q23.2	80,039,621	ASN	WWOX, MAF	Lu et al. 2019	30529582
rs1078643	17p12	10,707,241	ASN	PIRT	Lu et al. 2019	30529582
rs73975588	17p13.3	816,741	ASN	NXN	Law et al. 2019	31089142
rs9797885	19q13.2	41,873,001	ASN	TMEM91	Law et al. 2019	31089142
rs6055286	20p12.3	7,718,045	ASN	x	Law et al. 2019	31089142
rs2423279	20p12.3	7,812,350	ASN	HAO1	Jia et al. 2008	23263487
rs2179593	20q13.12	42,660,286	ASN	TOX2	Law et al. 2019	31089142
rs13831	20q13.32	57,475,191	ASN	GNAS	Lu et al. 2019	30529582
rs61776719	1p34.3	38,461,319	EUR	x	Law et al. 2019	31089142
rs10911251	1q25.3	183,112,059	EUR	LAMC1	Peters et al. 2013, Whiffin et al. 2014	23266556, 24737748
rs6687758	1q41	222,164,948	EUR	x	Houlston et al. 2010	20972440
rs6691170	1q41	222,045,446	EUR	DUSP10	Houlston et al. 2010	20972440

rs11692435	2q11.2	98,275,354	EUR	ACTR1B	Law et al. 2019	31089142
rs11893063	2q33.1	199,601,925	EUR	LOC105373831	Law et al. 2019	31089142
rs7593422	2q33.1	200,131,695	EUR	x	Law et al. 2019	31089142
rs992157	2q35	219,154,781	EUR	TMBIM1	Orlando et al. 2016	27005424
rs9831861	3p21.1	53,088,285	EUR	x	Law et al. 2019	31089142
rs12635946	3q13.2	112,916,918	EUR	x	Law et al. 2019	31089142
rs10936599	3q26.2	169,774,313	EUR	MYNN	Houlston et al. 2010	20972440
rs1370821	4q22.2	94,943,383	EUR	x	Schmit et al. 2018	29917119
rs17035289	4q24	106,048,291	EUR	x	Law et al. 2019	31089142
rs75686861	4q31.21	145,621,328	EUR	HHIP	Law et al. 2019	31089142
rs35509282	4q32.2	163,333,405	EUR	FSTL5	Schmit et al. 2014	25023989
rs58791712	5p13.1	40,281,797	EUR	PTGER4	Schmit et al. 2018	29917119
rs2735940	5p15.33	1,296,486	EUR	TERT	Schmit et al. 2018	29917119
rs2853668	5p15.33	1299910	EUR	TERT	Peters et al. 2012	21761138
rs62404968	6p12.1	55,714,314	EUR	BMP5	Schmit et al. 2018	29917119
rs1321311	6p21.2	36,622,900	EUR	CDKN1A	Dunlop et al. 2012	22634755
rs9271770	6p21.32	32,594,248	EUR	LOC107987449	Law et al. 2019	31089142
rs3131043	6p21.33	30,758,466	EUR	HCG20	Law et al. 2019	31089142
rs2070699	6p24.1	12,292,772	EUR	EDN1	Law et al. 2019	31089142
rs6928864	6q21	105,966,894	EUR	x	Law et al. 2019	31089142
rs10951878	7p12.3	46,926,695	EUR	x	Law et al. 2019	31089142
rs3801081	7p12.3	47,511,161	EUR	TNS3	Law et al. 2019	31089142
rs16892766	8q23.3	117,630,683	EUR	EIF3H	Tomlinson et al. 2008	18372905
rs1412834	9p21.3	22,110,131	EUR	CDKN2B-AS1	Law et al. 2019	31089142
rs10994860	10q11.23	52,645,424	EUR	A1CF	Schmit et al. 2018	29917119
rs10904849	10p13	16,955,267	EUR	CUBN	Al Tassan et al. 2015	25990418
rs4450168	11p15.4	10,286,755	EUR	SBF2	Law et al. 2019	31089142
rs3824999	11q13.4	74,345,550	EUR	POLD3	Dunlop et al. 2012	22634755
rs3802842	11q23.1	111,171,709	EUR	COLCA2	Tenesa et al. 2008	18372901
rs7136702	12q13.12	50,880,216	EUR	LARP4	Houlston et al. 2010	20972440
rs7398375	12q13.3	57,540,848	EUR	LRP1	Law et al. 2019	31089142
rs72013726	12q24.21	115,890,835	EUR	MED13L	Schmit et al. 2018	29917119
rs10161980	13q13.2	34,093,518	EUR	STARD13	Schmit et al. 2018	29917119
rs12427600	13q13.3	37,460,648	EUR	SMAD9	Law et al. 2019	31089142
rs45597035	13q22.1	73,649,152	EUR	KLF5	Law et al. 2019	31089142

rs1330889	13q22.3	78,609,615	EUR	LINC00446	Law et al. 2019	31089142
rs7993934	13q34	111,074,915	EUR	COL4A2	Law et al. 2019	31089142
rs1957636	14q22.2	54,560,018	EUR	BMP4	Tomlinson et al. 2011	21655089
rs4444235	14q22.2	54,410,919	EUR	BMP4	Houlston et al. 2008, Tomlinson et al. 2011	19011631, 21655089
rs11632715	15q13.3	33,004,247	EUR	x	Tomlinson et al. 2008	18372905
rs16969681	15q13.3	32,993,111	EUR	SCG5	Tomlinson et al. 2008, 2011	18372905, 21655089
rs4779584	15q13.3	32,994,756	EUR	CRAC1	Tomlinson et al. 2008	18372905
rs4776316	15q22.31	67,007,813	EUR	SMAD6	Law et al. 2019	31089142
rs10152518	15q23	68,177,162	EUR	x	Law et al. 2019	31089142
rs7495132	15q26.1	91,172,901	EUR	CRTC3	Law et al. 2019	31089142
rs9929218	16q22.1	68,820,946	EUR	CDH1	Houlston et al. 2008	19011631
rs61336918	16q23.2	80,007,266	EUR	x	Law et al. 2019	31089142
rs2696839	16q24.1	86,340,448	EUR	FOXF1	Schmit et al. 2018	29917119
rs4939827	18q21	46,453,463	EUR	SMAD7	Broderick et al. 2007, Tenesa et al. 2008	17934461, 18372901
rs285245	19p13.11	16,420,817	EUR	x	Law et al. 2019	31089142
rs10411210	19q13.11	33,532,300	EUR	RHPN2	Houlston et al. 2008	19011631
rs12979278	19q13.33	49,218,602	EUR	MAMSTR	Law et al. 2019	31089142
rs4813802	20p12.3	6,699,595	EUR	BMP2	Tomlinson et al. 2011, Peters et al. 2012	21655089 21761138
rs961253	20p12.3	6,404,281	EUR	BMP2	Houlston et al. 2008	19011631
rs2295444	20q11.22	33,173,883	EUR	PIGU	Schmit et al. 2018	29917119
rs1810502	20q13.13	49,057,488	EUR	PTPN1	Schmit et al. 2018	29917119
rs3787089	20q13.33	62,316,630	EUR	RTEL1	Law et al. 2019	31089142
rs4925386	20q13.33	60,921,044	EUR	LAMA5	Houlston et al. 2010	20972440
rs8124813	3p14.1	43,476,841	EUR, ASN	LRIG1	Schumacher et al. 2015	26151821
rs35360328	3p22.1	40,924,962	EUR, ASN	CTNNB1	Schumacher et al. 2015	26151821
rs1476570	6p22.1	29,809,860	EUR, ASN	HLA-G	Lu et al. 2019	30529582
rs2450115	8q23.3	117,624,093	EUR, ASN	EIF3H	Tomlinson et al. 2008	18372905
rs6469656	8q23.3	117,647,788	EUR, ASN	EIF3H	Tomlinson et al. 2008	18372905
rs6983267	8q24.21	128,413,305	EUR, ASN	POU5F1B	Haiman et al. 2007, Tomlinson et al. 2007, Hutter et al. 2010, Cui et al.2011	17618282, 17618284, 21129217, 21242260

rs12255141	10q25.2	114,294,892	EUR, ASN	VTI1A	Law et al. 2019	31089142
rs704017	10q22.3	80,819,132	EUR, ASN	ZMIZ1	Zhang et al. 2014	24836286
rs1035209	10q24.2	101,345,366	EUR, ASN	SLC25A28	Whiffin et al. 2014	24737748
rs4919687	10q24.32	104,595,248	EUR, ASN	CYP17A1	Zeng et al. 2016	26965516
rs11196172	10q25.2	114,726,843	EUR, ASN	TCF7L2	Zhang et al. 2014	24836286
rs12241008	10q25.2	114,280,702	EUR, ASN	VTI1A	Wang et al. 2014	25105248
rs1535	11q12.2	61,597,972	EUR, ASN	FADS2	Zhang et al. 2014	24836286
rs174550	11q12.2	61,571,478	EUR, ASN	FADS1	Zhang et al. 2014	24836286
rs4246215	11q12.2	61,564,299	EUR, ASN	FEN1	Zhang et al. 2014	24836286
rs174537	11q12.2	61,552,680	EUR, ASN	MYRF	Zhang et al. 2014	24836286
rs10849438	12p13.31	6,412,036	EUR, ASN	x	Law et al. 2019	31089142
rs10849432	12p13.31	6,385,727	EUR, ASN	PLEKHG6	Zhang et al. 2014	24836286
rs3217810	12p13.32	4,388,271	EUR, ASN	CCND2	Peters et al. 2013, Whiffin et al. 2014	23266556, 24737748
rs10774214	12p13.32	4,368,352	EUR, ASN	CCND2	Jia et al. 2013	23263487
rs12603526	17p13.3	800,593	EUR, ASN	NXN	Zhang et al. 2014	24836286
rs7229639	18q21.1	46,450,976	EUR, ASN	SMAD7	Zhang et al. 2014	24836286
rs1800469	19q13.2	41,860,296	EUR, ASN	TMEM91	Zhang et al. 2014	24836286
rs2241714	19q13.2	41,869,392	EUR, ASN	B9D2	Zhang et al. 2014	24836286
rs606682520	20q13.13	897,353	EUR, ASN	PREX1	Schumacher et al. 2015	26498495
rs6066825	20q13.13	47,340,117	EUR, ASN	PREX1	Schumacher et al. 2015	26151821
rs5934683	Xp22.2	9,751,474	EUR, ASN	SHROOM2	Dunlop et al. 2012	22634755

*Abbreviation: *CHR,* chromosome; *POS,* position; *POP,* population (AFR A – African American, ASN – Asian, EUR – European).

Tab. 2 Outliers identified in boxplots that show allele frequency differences of Slovak and the other six world populations for 106 risk CRC variants identified from GWAS.

rs_ID	CHR : POS	Variant type	GENE : CONSEQUENCE	CLINICAL SIGNIFICANCE	POPULATION COMPARISON
rs5934683	chrX:9783434	SNV	GPR143 : Intron Variant	Not Reported in ClinVar	Slovak-East Asian
rs7252505	chr19:33084158	SNV	GPATCH1 : Intron Variant	Not Reported in ClinVar	Slovak-African
rs4779584	chr15:32702555	SNV	None	Not Reported in ClinVar	Slovak-East Asian
rs174550	chr11:61804006	SNV	FADS1 : Intron Variant	Not Reported in ClinVar	Slovak-American
rs4246215	chr11:61796827	SNV	FEN1 : 3 Prime UTR Variant	Not Reported in ClinVar	Slovak-American
					Slovak-European (Finnish)
					Slovak-European (non-Finnish)
					Slovak-East Asian
rs10904849	chr10:16955267	SNV	CUBN : Intron Variant	Not Reported in ClinVar	Slovak-European (non-Finnish)
rs6928864	chr6:105519019	SNV	None	Not Reported in ClinVar	Slovak-African
rs3131043	chr6:30790689	SNV	HCG20 : Intron Variant	Not Reported in ClinVar	Slovak-European (non-Finnish)
					Slovak-European (Finnish)
rs12659017	chr5:126652483	SNV	None	Not Reported in ClinVar	Slovak-East Asian
rs397775554	chr5:40281696-40281704	Indel	None	Not Reported in ClinVar	Slovak-European (non-Finnish)

Tab. 3 18 selected variants (UTR and non-coding) from all 648 variants identified in genes associated with Lynch syndrome from NIPT data in the Slovak population.

rs_ID	CHR : POS	VARIANT TYPE	GENE	CONSEQUENCE	CLINICAL SIGNIFICANCE
rs11901645	chr2:47510079	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs11891189	chr2:47510259	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs876936	chr2:47513059	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs72872839	chr2:47565070	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs17036769	chr2:47633503	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs2969774	chr2:47661684	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs2952372	chr2:47661919	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs2969773	chr2:47662141	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs2705765	chr2:47662389	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs12328344	chr2:47662542	SNV	MSH2	3 prime UTR variant	Not reported in ClinVar
rs10427209	chr2:47709297	SNV	MSH6	Non-coding transcript exon variant	Not reported in ClinVar
rs10427344	chr2:47709476	SNV	MSH6	Non-coding transcript exon variant	Not reported in ClinVar
rs3136240	chr2:47784947	SNV	MSH6	Non-coding transcript exon variant	Not reported in ClinVar
rs6791557	chr3:30614676	SNV	TGFBR2	Non-coding transcript exon variant	Not reported in ClinVar
rs1817338	chr3:30631239	SNV	TGFBR2	Non-coding transcript exon variant	Not reported in ClinVar
rs9852378	chr3:36997280	SNV	MLH1	Non-coding transcript exon variant	benign in ClinVar
rs10951972	chr7:6002187	SNV	PMS2	Non-coding transcript exon variant	Not reported in ClinVar
rs10951973	chr7:6002205	SNV	PMS2	Non-coding transcript exon variant	Not reported in ClinVar

Tab. 4 15 pathogenic and likely pathogenic variants associated with Lynch syndrome with non-zero allele frequency in gnomAD database.

				ALLELE FREQUENCY IN POPULATION					
rs_ID	CHR : POS	REF Allele	ALT Allele	African	American	East Asian	European (Finnish)	European (non-Finnish)	South Asian
rs63750615	2:47403333	G	T	0	0	0	0	0	0.000328
rs1194793421	2:47414417	AG	A	0	0	0	0	2.75E-05	0
rs63750636	2:47476492	C	T	0	0	0	0	1.55E-05	0
rs63749873	2:47795903	C	G	0	0	0	0	3.10E-05	0
rs587783056	2:47799684	GTT	G	4.76E-05	0	0	0	0	0
rs63751017	2:47800714	C	T	0	0	0	0	3.10E-05	0
rs876660943	2:47806359	G	T	0	0	0	0	1.55E-05	0
rs63751221	3:37001045	C	T	2.38E-05	0	0	0	0	0
rs587779338	7:5977589	G	A	4.86E-05	0	0	0	1.58E-05	0
rs267608161	7:5982885	C	T	4.80E-05	0	0	9.61E-05	1.55E-05	0
rs63751422	7:5986838	G	A	2.38E-05	0	0	0	0	0
rs63750250	7:5986933	A	AT	0	0	0	0	9.30E-05	0
rs200640585	7:5992018	G	A	0	0	0	0	1.55E-05	0
rs267608154	7:5995572	ACTGT	A	0	0	0	0	1.55E-05	0
rs63750871	7:6002590	G	A	0	7.33E-05	0	0	0	0

Figures

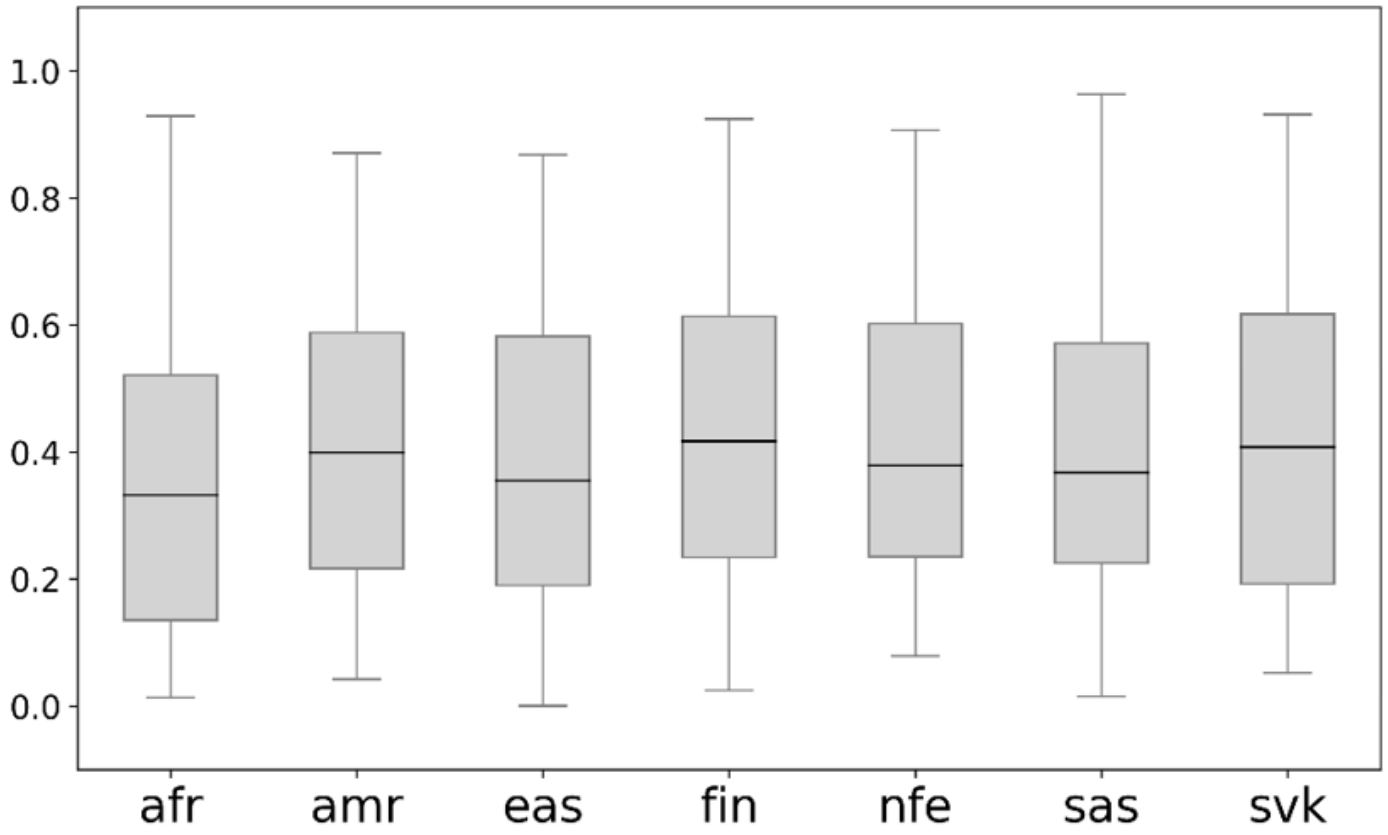

Figure 1

Boxplots show allele frequency of 106 risk CRC variants identified from GWAS for the Slovak and the other six world populations

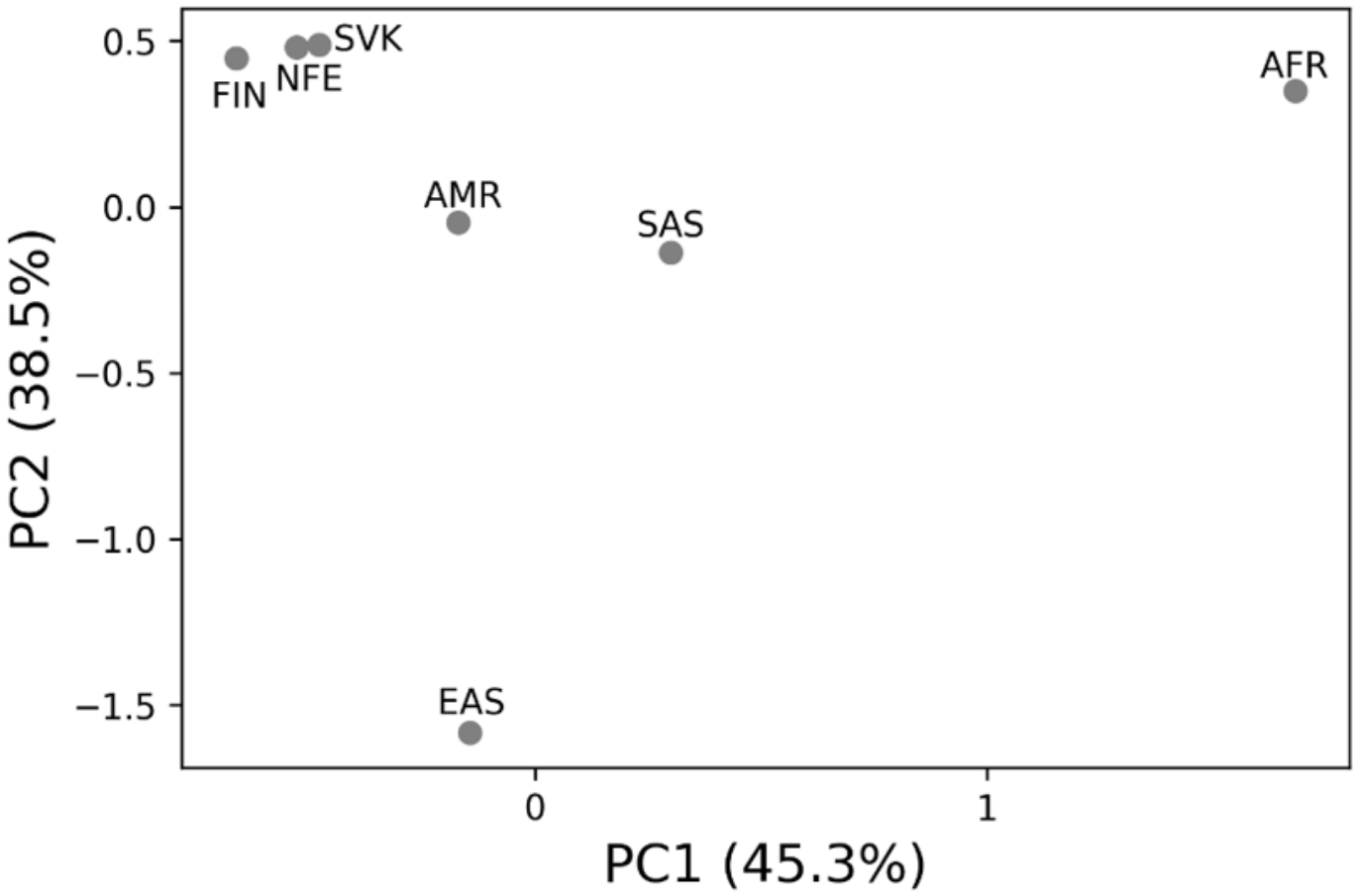

Figure 2

PCA plot illustrates the allele frequency of 106 risk CRC variants identified from GWAS for the Slovak and the other six world populations

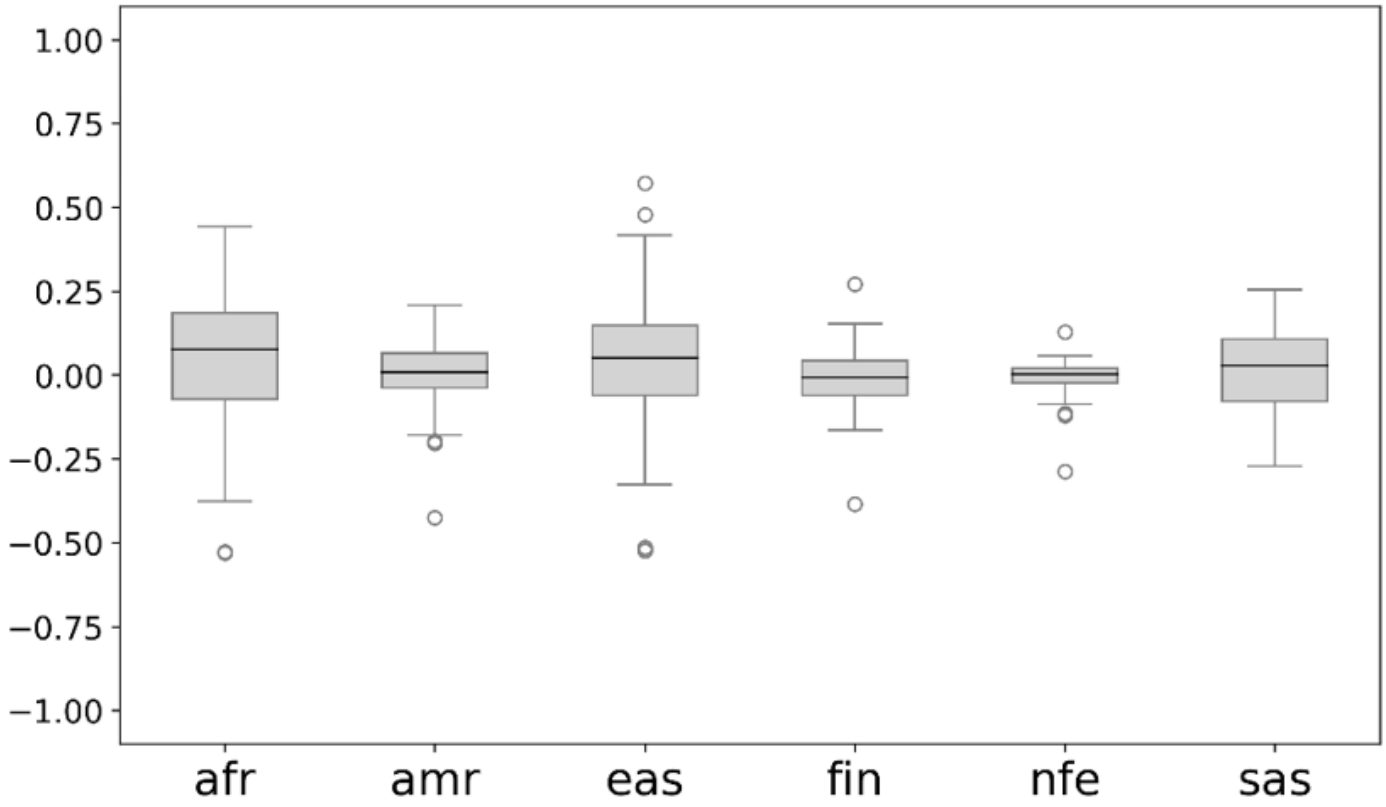

Figure 3

Boxplots show allele frequency differences of Slovak and the other six world populations for 106 risk CRC variants identified from GWAS

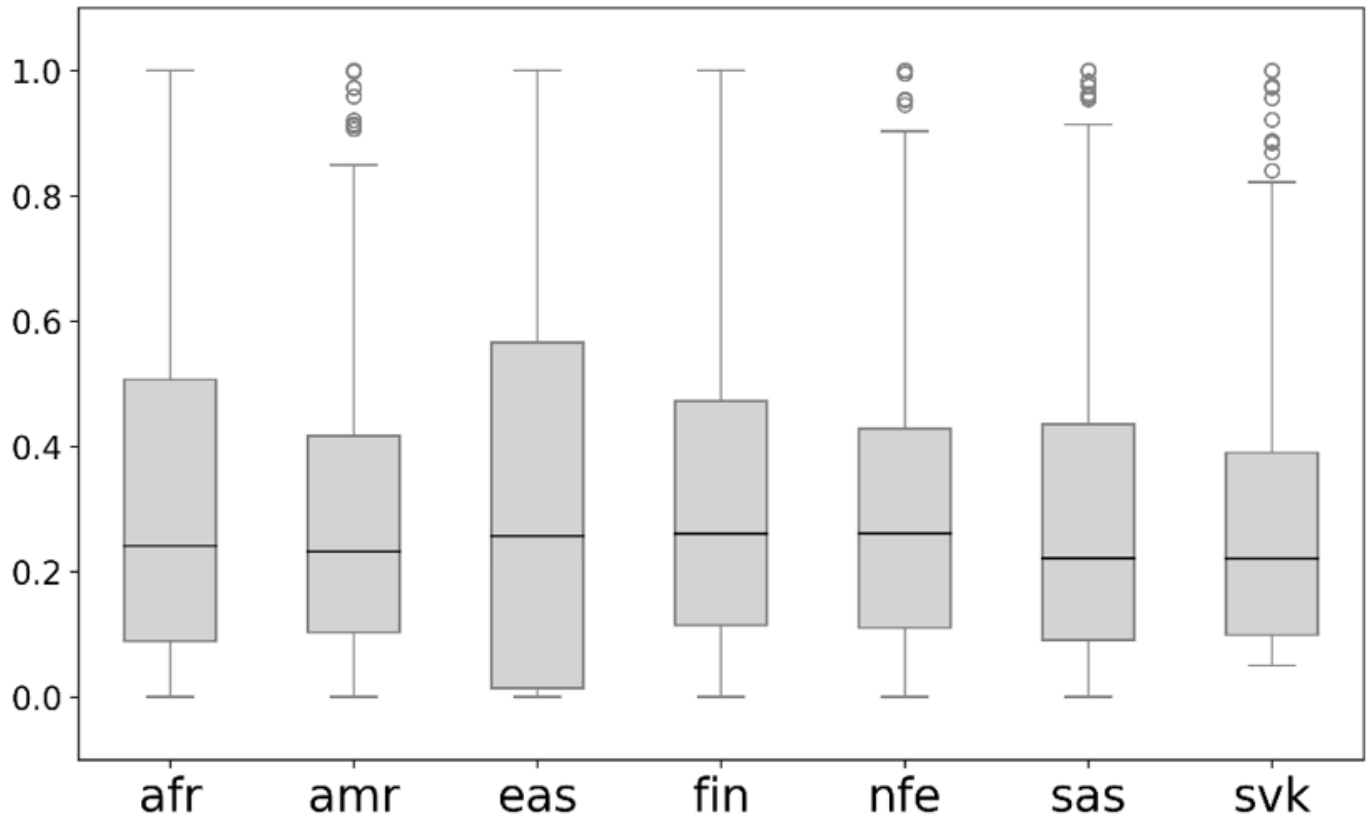

Figure 4

Boxplots show allele frequency of 648 variants located in 7 Lynch genes for the Slovak and the other six world populations

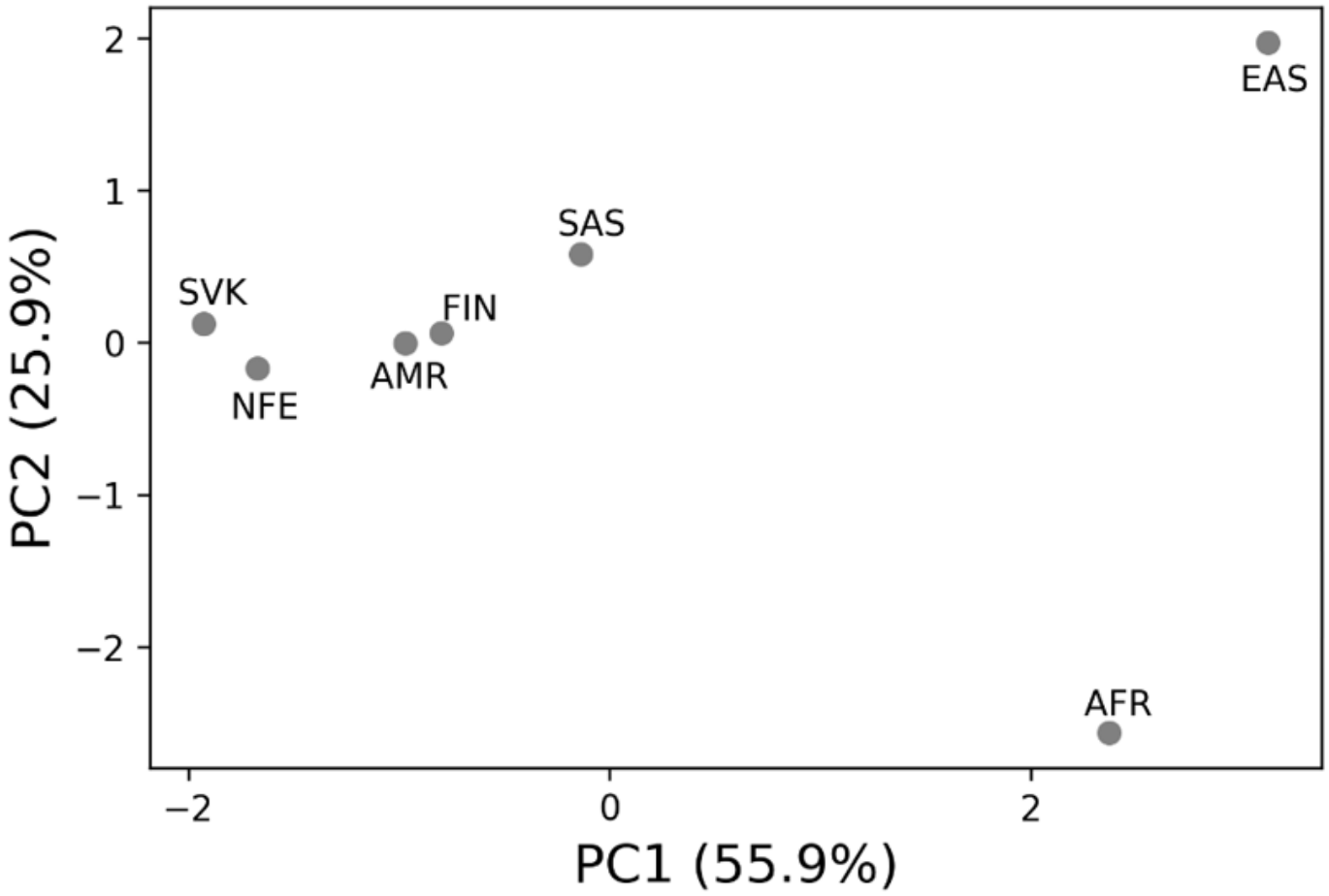

Figure 5

PCA plot illustrates the allele frequency of 648 variants located in genes associated with LS for the Slovak and the other six world populations

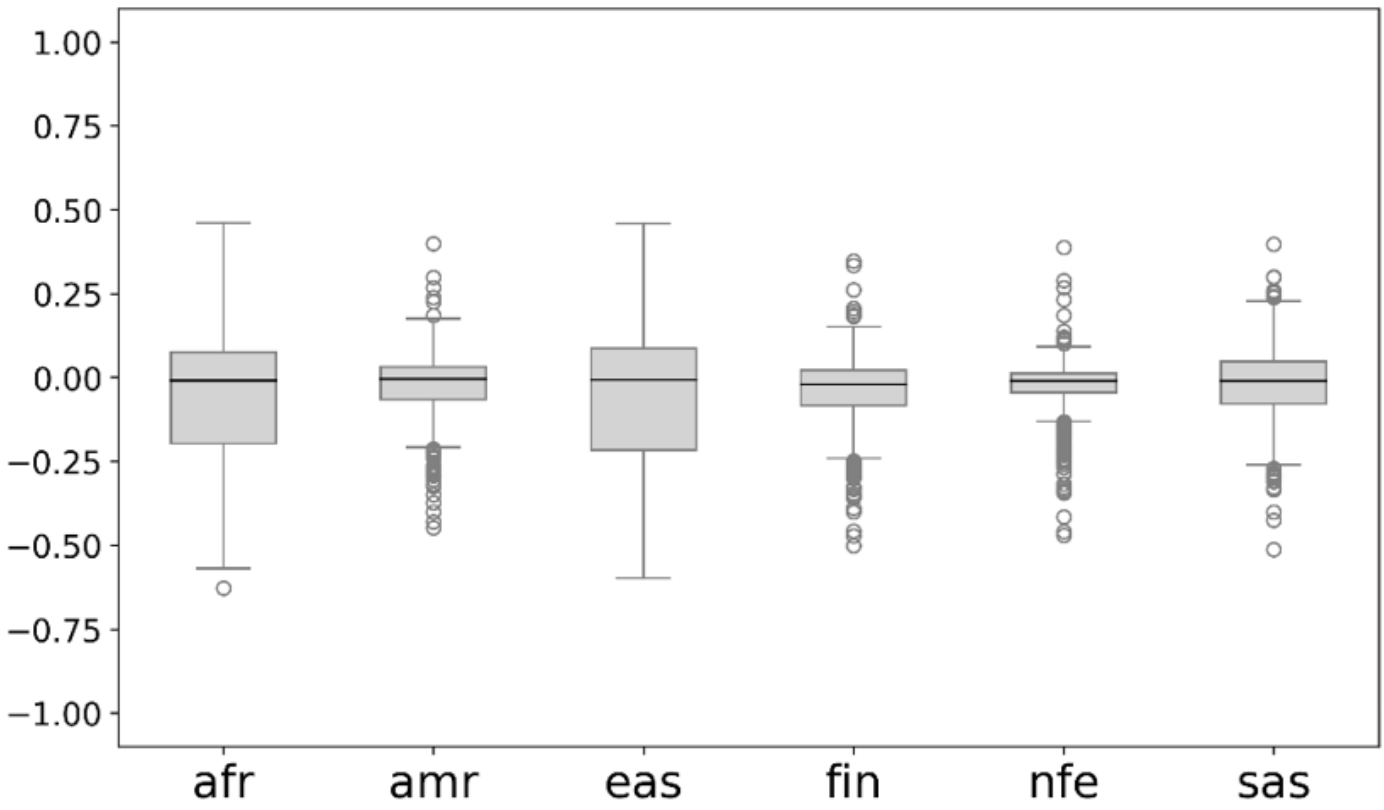

Figure 6

Boxplots show differences of Slovak and the other six world populations in allele frequency for 648 variants located in genes associated with LS

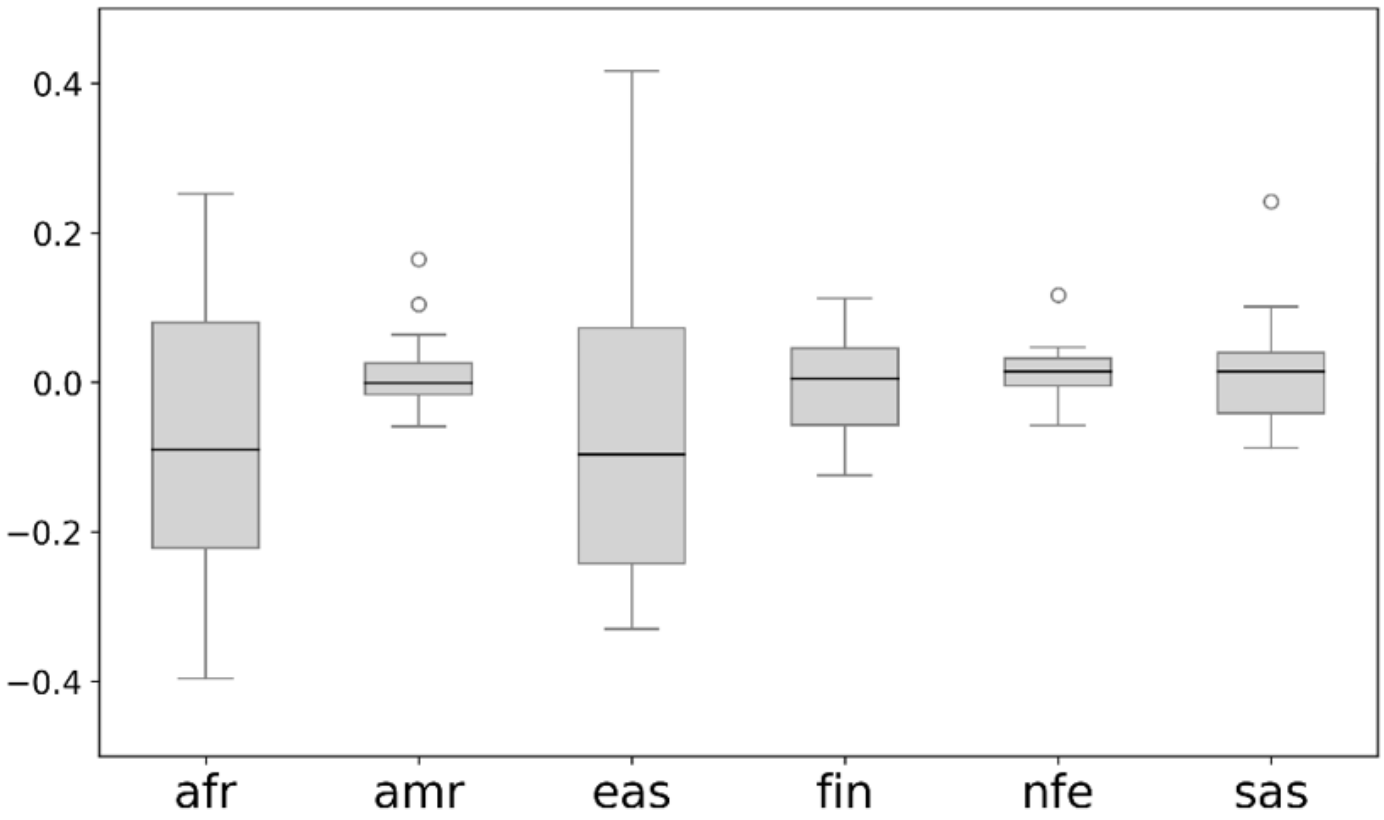

Figure 7

Boxplots show differences of Slovak and the other six gnomAD world populations in allele frequency for 18 selected variants (UTR and non-coding variants) from 648 identified variants located in risk LS genes

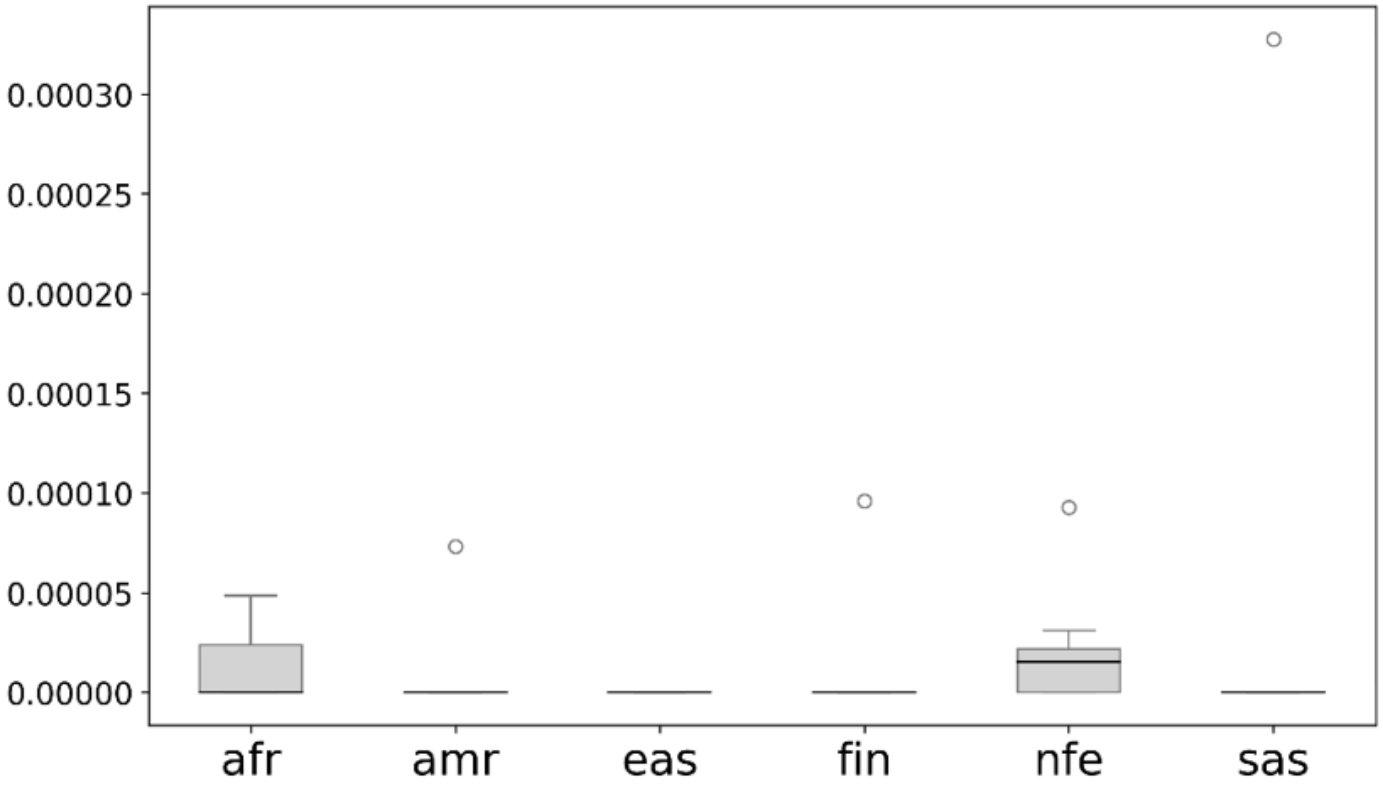

Figure 8

Boxplots show allele frequency of 15 SNPs of LS genes with pathogenic and likely pathogenic clinical significance for six world populations. All found allele frequency are highly below 5%