

UDC 575-004.9

OPTIMIZATION OF GENOME-WIDE SNP MICROARRAY DATA PROCESSING FOR GENERATION OF AN ATHLETE GENETIC PROFILE

T.N. Kereyev¹, A.V. Mustafayeva^{1,2*}, Zh.A. Zhaniyazov^{1,2,3}, N.K. Altynova¹, G.S. Zhunussova¹, L.B. Djansugurova^{1,2*}, N. M. Kassymbek², S. T. Mukhambetzhano²

¹Institute of Genetics and Physiology, CS MSHE RK, Almaty, Kazakhstan

²Al-Farabi Kazakh National University, Almaty, Kazakhstan

³Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan

*e-mail: mustafayeva_a@live.kaznu.kz (A.V.); leylad@mail.ru (L.B.)

ABSTRACT

Based on the analysis of global experience and our own preliminary studies aimed at identifying associations between athletic performance potential and individual genotypes across 97 SNP polymorphisms (75 candidate genes), interpretations of possible genotype variants for an athlete's genetic passport were developed. A software program and several auxiliary scripts were created using the RStudio development environment and the Python programming language. The developed tools efficiently create and update databases containing the required genetic information, reducing the time needed to develop a genetic passport while minimizing the risk of errors. These tools significantly improved the efficiency of storing, sorting, and processing primary data obtained from genome-wide SNP microarray genotyping. The reduction of human involvement in the repetitive process of collecting and comparing tables minimized the risk of errors and facilitated the analysis of genotyping data. Automation of primary and secondary processing of genome-wide SNP microarray genotyping data, genotype interpretation, database updating, and generation of completed genetic passport reduced data processing time from several months to 20-30 minutes.

Keywords: sports genetics, genetic passport, genome-wide SNP microarray genotyping, single nucleotide polymorphism (SNP), software code

INTRODUCTION

The rapid development of modern genotyping technologies has greatly improved access to information regarding both population-wide and individual genetic variation in humans depending on genetic polymorphism. Due to the achievements of GWAS (Genome-Wide Association Studies), it has become clear that common human diseases and complex physical traits are determined not by single mutations but by the cumulative contribution of many genetic variants resulting from single-nucleotide polymorphisms (SNPs). Contemporary approaches, including whole-genome sequencing and genome-wide SNPs-microarray analysis, make it possible to obtain extensive genomic information from relatively small amounts of biological material. These technologies enable large-scale analysis of nuclear and mitochondrial genetic variation with high speed and accuracy.

The role of hereditary factors in determining physiological characteristics and athletic abilities has become an important topic in sports science research. Elite athletic performance is associated with the realization of an individual's physical potential, which is believed to be strongly influenced by genetic background. Current estimates suggest that hereditary factors account for approximately 75-80% of athletic predisposition, while factors such as coaching, training intensity, diet, lifestyle, and psychological preparedness further influence the development and expression of sport-specific abilities.

Despite being a relatively young field of research, sports genetics has already demonstrated considerable practical value. Its applications are particularly relevant during childhood and adolescence, when many physiological characteristics associated with future athletic specialization are established and developed. As a result, genetic information is increasingly being considered as a complementary tool for un-

derstanding individual differences in sport-related traits and physical capabilities. Recent genome-wide association studies have identified a large number of polymorphic loci associated with athletic performance and exercise-related traits [1, 2]. Some genetic variants are linked not only to enhanced performance in certain categories of sports but also to predisposition to sports-related injuries. Complex athletic traits such as muscular strength, endurance capacity, speed, recovery after physical activity, and adaptive muscle responses are controlled through interactions among multiple genes and environmental factors. The combined effects of these mechanisms contribute to the formation of important athletic characteristics including power, speed, and endurance [3,9].

Metabolic variability and hormonal regulation are also essential considerations in professional sports because athletes are regularly exposed to intense physical exercises. Genetic differences may influence susceptibility to injuries, exercise-induced disorders, and dysfunctions affecting particular organs or physiological systems [10]. In addition to physical performance, psychological stability plays a critical role in competitive sports. The demands of elite-level sports place considerable psychological and emotional strain on athletes. As a result, effective stress management and adaptive capacity are recognized as key factors influencing training outcomes and competitive success. Behavioral and neuropsychological traits such as motivation, stress tolerance, leadership qualities, decision-making ability, prioritization, and teamwork may substantially influence athletic success [11,12].

Within the framework of the targeted research project entitled "Development of an integrated system for training elite athletes and prospective Olympic reserves in sports prioritized in Kazakhstan based on physiological and genetic assessment" (BR18574139), a genetic passport for athletes was created. The proposed system provides a comprehensive as-

assessment of the influence of specific genotypes and allelic variants on inherited traits associated with athletic performance as well as susceptibility to sports-related disorders and pathological conditions.

The application of a genetic passport is not limited to the identification and selection of talented athletes. It also represents a valuable tool for designing personalized training programs that take into account the unique genetic and physiological characteristics of an individual. Furthermore, genetic information can be used to identify potential health risks, evaluate individual metabolic responses to nutrients and dietary components, and support the development of customized nutritional strategies based on genomic profiles.

Implementation of personalized genomic approaches in sports may support more effective decision-making regarding training processes, nutrition, and rehabilitation strategies, ultimately improving athletic performance and reducing the probability of injuries. Furthermore, individualized approaches based on genetic data may contribute to higher competitiveness in international sports. Nevertheless, the collection, storage, and interpretation of genetic information require careful consideration of ethical principles, confidentiality, and data protection.

Despite the high efficiency of modern technologies for generating genomic information, processing and interpretation of genome-wide SNP-array datasets remain labor-intensive and computationally demanding. Primary and secondary analysis of genome-wide data may require long processing times and frequently involve customized analytical procedures. Therefore, practical and commercial implementation of such technologies depends on optimization of bioinformatic pipelines and automation of analytical workflows. In studies aimed at rapid interpretation of genomic data, including evaluation of hereditary disorders, genetic susceptibility to diseases, or assessment of athletic potential, efficient filtering strategies and optimized computational approaches are especially important. Improved data processing methods also facilitate more secure management and storage of sensitive genetic information obtained from biological samples.

Optimization of genome-wide SNP-array data-processing workflows may accelerate the extraction of relevant genotype information and the generation of standardized individual reports. Such developments may contribute to more accurate identification of athletic predispositions, enhancement of training methodologies, and prevention of sports-related injuries.

Accordingly, the aim of the present study was to optimize the processing and storage of genome-wide SNP genotyping data and to improve the generation of athlete genetic passport using computational and bioinformatic approaches.

MATERIALS AND METHODS

Study subjects

The present study utilized previously generated genome-wide SNP microarray data comprising approximately 654,027 genetic markers from 1,415 participants. The dataset included three groups of participants. The first group comprised 934 professional athletes, including candidates for master of sport, masters of sport, champions, and prize winners of regional, national, Asian, and international competitions. The

second group consisted of 239 recreational athletes who regularly attended fitness centers and gyms, while the third group included 242 individuals who did not participate in organized sports activities. Questionnaire data and written informed consent had been obtained from all participants during the original data-collection stage (research project BR18574139 “Development of an integrated system for training highly qualified athletes and promising Olympic reserves in sports prioritized in Kazakhstan based on physiological and genetic assessment” for the 2022-2024). The study protocol was approved by the Local Ethics Committee of the Institute of Genetics and Physiology under the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (*Protocol No. 5, July 25, 2022*). Confidentiality of personal and genetic information was ensured throughout the study, and participants retained the right to control disclosure and interpretation of results related to their genetic characteristics and athletic predisposition.

Our study objects included DNA samples from provided collection, raw genotyping individual data obtained with Illumina iScan platform were our study objects, that processed in the present study. Questionnaire, physical, physiological, and biochemical data collected within the previous project were not reanalyzed. Previously established associations between genotypes and physical and biochemical characteristics were used as the basis for developing the interpretation framework for the athlete genetic profile [13].

Genome-wide SNP microarray genotyping

Genome-wide SNP microarray genotyping was performed using the Illumina iScan System and the Infinium Global Screening Array-24 v3.0 Kit (Illumina, USA), which assays approximately 654,027 predefined genetic markers distributed across the genome.

DNA samples with a minimum concentration of 50 ng/μL and a purity ratio of 1.75–1.80 were used for SNP microarray genotyping. Sample preparation was performed using the Infinium Automation Kit – 8 Tip Tecan Non-LIMS automated workstation (Illumina, USA). Genotyping was carried out according to the Infinium HTS Automated Workflow.

Processing of genome-wide SNP genotyping data

Initial processing and quality assessment of the SNP microarray data were performed using Illumina GenomeStudio v.2010.3 (Illumina, USA). Samples with genotype call rates below 95% were excluded from further processing. The resulting genotype data were exported in PLINK-compatible formats for subsequent extraction and structuring.

The bioinformatic workflow included verification of sample identifiers, SNP identifiers, chromosomal and mitochondrial localization of the markers, genomic coordinates, and homozygous or heterozygous genotype states. Genotype calls were represented as diploid genotypes for each biallelic SNP.

Annotation of genetic variants

Candidate genetic variants were annotated using the Ensembl BioMart package [14, 15]. BioMart was used to retrieve information on SNP identifiers, chromosomal locations, genomic coordinates, corresponding genes, and functional annotations. All genomic coordinates and variant annotations used in the present study corresponded to the GRCh38/hg38 reference genome assembly. Additional information on previously

reported SNP–phenotype associations was obtained from publicly available genomic resources, including the GWAS Catalog (<https://www.ebi.ac.uk/gwas/>), dbSNP (<https://www.ncbi.nlm.nih.gov/snp/>), ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), and the 1000 Genomes Project database (<https://www.internationalgenome.org/1000-genomes-browsers>).

Information on previously established associations between genetic variants and physical and biochemical characteristics of athletes was obtained from the report of the research project “Development of an integrated system for training highly qualified athletes and promising Olympic reserves in sports prioritized in Kazakhstan based on physiological and genetic assessment” (BR18574139-OT-23).

Based on the BioMart annotations, information from public genomic databases, and previously established genotype–phenotype associations, a panel of 97 SNPs located in or near 75 candidate genes was compiled. For each selected SNP, possible diploid genotypes, their corresponding interpretations, and predefined reporting scores were included in the interpretation database [13]. The list of selected SNPs is represented as Supplementary Table S1.

Structuring of genome-wide SNP genotyping data

Individual-level SNP microarray genotype datasets were extracted using PLINK v1.9. Data-extraction procedures were automated using Python version 3.13.3. Subsequent sorting and organization of the extracted genotype information were performed in R, with RStudio version 2024.04.2+764 used as the development environment.

The 97 SNPs were organized into four reporting categories: physical performance traits, predisposition to sports-related injuries and pathological conditions, metabolism and nutrition, and psychogenetic characteristics. The annotated variants, genotype-specific interpretations, and predefined reporting scores were stored in Microsoft Excel (.xlsx) format.

Scaling and scoring the phenotype expression

To facilitate interpretation of the results by athletes and qualified specialists, a standardized 10-point scale was applied separately to each evaluated characteristic. We've been supposed that the maximum score of 10 points represents the highest interpreted expression of the corresponding phenotype or low risk of sport-related disease development. Score of 5 point corresponds to typical phenotype/risk expression (average or norm). Scores below 5 represented a low level of athlete's phenotype expressivity or high risk of disease development.

Each characteristic responsible genotype interpretation was evaluated regarding the trait expressivity.

When a characteristic was represented by multiple SNPs, the predefined scores assigned to the observed genotypes scores were summed according to the following equation:

$$P = \sum(x_1 + x_2 + \dots + x_n), \quad (1)$$

where P is the final score for the evaluated characteristic (maximum 10 points), is the number of SNPs included in the assessment. For each characteristic, genotype-specific values were assigned so that the sum of the maximum possible contributions of all included SNPs was equal to 10 points. The maximum point for each SNP input was estimated considering the function of related genes. So, if the individual presents

the favorable genotypes for all responsible SNPs, the score of trait expressivity should give the maximum score (10 points).

Databases and storage of structured data

Processed genotyping information together with genotype interpretation results was stored in both .txt and Microsoft Excel (.xlsx) file formats for further analysis and integration into the genetic passport system. The interpretation database contained the selected SNPs, genotype-specific interpretations, and predefined reporting scores.

Generation of the final genetic passport

Generation of the final athlete genetic passport was automated using R and Python scripts. The R based scripts collected and organized genotype-specific information for each participant, whereas the Python scripts automatically assembled the final report in Microsoft Word format.

For each participant, the program retrieved the observed genotypes for the selected SNPs and matched them with the corresponding genotype-specific interpretations and predefined reporting scores stored in the interpretation database. The resulting information was organized according to the four reporting categories and transferred to the final standardized document.

RESULTS AND DISCUSSION

Structure of the genetic passport and formation of the interpretation database

Interpretations of individual genotypes were based on previous project studies, relevant scientific publications [2, 4, 7, 8, 10], and information obtained from public genomic resources, including the GWAS Catalog, dbSNP, ClinVar, Ensembl, and the 1000 Genomes Project. According to the final report of the research project “Development of an integrated system for training highly qualified athletes and promising Olympic reserves in sports prioritized in Kazakhstan based on physiological and genetic assessment” (BR18574139-OT-23), an athlete genetic passport comprising a panel of 97 SNPs associated with 75 candidate genes was developed [13]. The selected SNPs were associated with physical performance, physiological adaptation, metabolism and nutrition, susceptibility to sports-related injuries and pathological conditions, and psychogenetic characteristics. The 97 SNPs were organized into four major reporting categories:

Physical performance traits, represented by 44 SNPs associated with strength, power, speed, endurance, muscle-fiber characteristics, and other traits relevant to physical performance;

Predisposition to sports-related injuries and pathological conditions, represented by 34 SNPs associated with injury susceptibility, tissue-repair processes, and adverse responses to intensive physical activity;

Metabolic and nutritional characteristics, represented by 38 SNPs associated with energy metabolism, micronutrient utilization, body-composition regulation, and nutritional characteristics;

Psychogenetic characteristics, represented by 30 SNPs associated with behavioral, cognitive, and psychological traits, including stress response, motivation, decision-making, and risk-taking behavior.

	A	B	C	D	E	F
82	Мышечная сила					
83	Ген рецептора активина A	ACVR1B A>C, G	rs2854464	AA	Хорошая мышечная сила	5
84				AC	Хорошая мышечная сила	5
85				CC	Мышечная сила не выражена	0
86				AG	Хорошая мышечная сила	5
87				GG	Мышечная сила не выражена	0
88		FTO	rs9939609	TT	Нормальный контроль приема пищи. Обычная сила мышц.	1
89				TA	Слегка повышенное потребление углеводов и высококалорийной пищи как у детей, так и у взрослых. Сила мышц увеличена.	3
90				AA	Повышенное потребление углеводов и высококалорийной пищи как у детей, так и у взрослых. Хорошая мышечная сила.	5
91	Выносливость при недостатке кислорода					
92	C>T	HIF1A	rs11549465	CC	Синтез фактора гипоксии HIF1-A, обеспечивающего нормальную жизнедеятельность клеток в условиях недостатка	1
93	Ген фактора гипоксии HIF1-A			CT	Генотип определяет легкое увеличение синтеза фактора гипоксии HIF1-A, обеспечивающего нормальную жизнедеятельность клеток в условиях недостатка кислорода. Это способствует адаптации спортсменов к длительным	1,5
94				TT	Генотип предпологает к увеличению синтеза фактора гипоксии HIF1-A, обеспечивающего нормальную жизнедеятельность клеток в условиях недостатка кислорода (гипоксии). Способствует лучшей адаптации спортсменов к длительным анаэробным нагрузкам.	2
95		ADRB2 Gly16Arg G>A, C	rs1042713	GG	Генотип по гену адренергического β2-рецептора не способствует развитию выносливости при анаэробных тренировках. Является фактором	0

Figure 1. Image of interpretation database source Excel file

Individual SNPs could be included in more than one trait or reporting category. Therefore, the category-specific SNP counts were not mutually exclusive, and their sum exceeded the total number of 97 unique markers.

Based on functional annotations and previously reported genotype–phenotype associations, genotype-specific interpretations were developed for inclusion in the athlete genetic passport. These interpretations were designed to present genetic information in a standardized format suitable for qualified sports physicians, coaches, and athletes.

All evaluated characteristic were included in specially created an interpretation database using Microsoft Excel format. The database consists from 4 worksheets according 4 structured major categories of trait sets: physical performance traits; predisposition to sports-related injuries and pathological conditions; metabolic and nutritional characteristics; psychogenetic characteristics. The Figure 1 represents the structure of interpretation database. Within each worksheet, data are organized on a trait-specific basis, where each athletic characteristic is described independently. A separate sheet provides the trait designation, a curated panel of SNP markers linked to the trait, corresponding genotype profiles, functional interpretation of observed genotype variants, descriptions of expected phenotypic outcomes, and a numerical assessment on a 10-point scale.

Athletic characteristics may involve analysis of a single gene or multiple genes together with corresponding SNP polymorphisms [7, 8]. In cases where the athletic characteristics includes multiple polymorphisms across several genes SNPs, the final interpretation of trait expressivity was also predicted.

To facilitate interpretation of genotype result by athletes,

we applied the scoring system supported by images for visual explanation. A standardized 10-point scale was applied separately to each evaluated characteristic (see Materials and Methods part). It has been supposed that the maximum score of 10 points represents the highest interpreted expression of the corresponding phenotype or low risk of sport-related disease development. Five points corresponds to an average or normative level of the characteristic expression. Taking into account the number of related genes (SNPs) and their impact to trait expressivity, the final maximum score of each physical or biochemical characteristics should be summarized as 10 maximum points.

As an example, the muscle-strength characteristic was evaluated using rs2854464 in or near *ACVR1B* and rs9939609 in *FTO*. Each marker had a maximum predefined contribution of 5 points, resulting in a maximum combined score of 10 points.

For *ACVR1B* rs2854464, the corresponding genotype groups were assigned predefined values of 5 or 0 points according to their interpreted relative contribution to the muscle-strength characteristic. For *FTO* rs9939609, genotypes AA, AT, and TT were assigned 5, 3, and 1 points, respectively.

Muscle fiber type expressivity depends on function of 4 genes – *PPARGC1A* (rs8192678), *PPARG* (rs 1801282), *PPARA* (rs 4253778), *TTN* (rs 10497520). *PPARGC1A* encodes transcriptional coactivator that *regulates the genes involved in energy metabolism*. The protein encoded by *PPARG* gene is a ligand-activated nuclear receptor and master regulator of adipogenesis that drives transcriptional programs for adipocyte differentiation, lipid storage, and glucose homeostasis. *PPARA* encodes peroxisome proliferator-activated receptor alpha, that links lipid sensing to gene regulation. It functions as

Speed-strength activities, Muscle hypertrophy, Endurance	ACTN2	rs223759	CC	Large muscle mass gain, excellent endurance. A significant advantage in various sports.	6
			CT	Moderate muscle mass gain and endurance. You have the potential to achieve high performance.	3
			TT	Low muscle mass gain, good endurance during physical activity, and a low body mass index (BMI).	0
	AMPD1	rs17620706	CC	The genotype determines normal activity of the AMP deaminase enzyme. No rapid onset of muscle fatigue.	4
			CT	The genotype determines reduced activity of the AMP deaminase enzyme (see mutant allele).	2
			TT	The genotype contains two mutant alleles of the AMP deaminase gene, resulting in a complete absence of AMP deaminase enzyme activity.	0
Muscle fiber types, endurance	PPARGC1A	rs652676	AA	The genotype determines normal utilization of fatty acids and glucose, which is associated with high endurance.	3
			GA	The genotype determines normal utilization of fatty acids and glucose, which is associated with moderate endurance.	2
			AA	The genotype determines reduced utilization of fatty acids and is associated with obesity in men.	1
	PPARG	rs1771287	CC	Normal body mass index (BMI). Effective weight loss in response to a calorie-restricted diet.	3
			CG	Increased body mass index (BMI). A tendency toward less effective weight loss in response to a calorie-restricted diet.	2
			GG	High body mass index (BMI). Poor weight loss response to a calorie-restricted diet.	1
	PPARG	rs4554776	AA	Good endurance. Better weight loss results through fitness training, particularly aerobic exercises.	3
			AG	Reduced fat and carbohydrate metabolism, associated with significantly greater progression of coronary artery disease.	2
			CC	Reduced fat and carbohydrate metabolism, associated with significantly greater progression of coronary artery disease.	1
	TTN	rs22487520	TT	Shorter skeletal muscle fascicle length. Associated with high personal performance in the sprint discipline.	1
			TC	Increased skeletal muscle fascicle length. Associated with good personal performance in the sprint discipline.	1
			CC	Longer skeletal muscle fascicles. Theoretically, this ensures that a smaller proportion of functional muscle fibers is attached during contraction, improving muscle	0
Muscular strength	ACTR3B	rs2054464	AA	Good muscular strength.	3
			AG	Good muscular strength.	2
			GG	Muscular strength is not pronounced.	0
	FTO	rs4556626	TT	Normal regulation of food intake. Normal muscle strength.	1
			TA	Slightly increased consumption of carbohydrates and high-calorie foods in both children and adults.	2
			AA	Increased consumption of carbohydrates and high-calorie foods in both children and adults.	3
Endurance under oxygen deficient conditions	HIF1A	rs22344440	CC	The genotype determines normal synthesis of the hypoxia-inducible factor HIF-1α and VEGF, ensuring normal cellular function under low oxygen conditions.	3
			CT	The genotype determines a slight increase in the synthesis of the hypoxia-inducible factor HIF-1α and VEGF, ensuring normal cellular function under low oxygen conditions.	2
			TT	The genotype predisposes to increased synthesis of the hypoxia-inducible factor HIF-1α and VEGF, ensuring normal cellular function under low oxygen conditions.	1
	ACORD2	rs2240712	GG	The genotype of the 50-ohm high-resistance (HMR) gene does not promote the development of endurance during physical activity.	0

Figure 2. Database of athletic characteristic interpretations with scoring assessment

a transcriptional regulator in the nucleus, where it heterodimerizes with *RXRα* to control metabolic programs. This receptor is central to fasting-associated adaptation and coordinates pathways involved in fatty acid oxidation, ketogenesis, and peroxisomal beta-oxidation. *TTN* encodes titin, which is the largest known protein in the human body and acts as a giant molecular spring in heart and skeletal muscles. The genes *PPARGC1A*, *PPARG*, *PPARA*, and *TTN* play critical roles in defining skeletal muscle fiber types, driving the shift between oxidative (Type I, slow-twitch) and glycolytic (Type II, fast-twitch) profiles [9, 16]. The molecular mechanisms underlying the impact of muscle fiber types were explained in animals [17]. The score for each gene impact was estimated taking into account the function of responsible genes and concrete SNP association with trait expressivity (figure 2).

Organizing information in this manner enables automatic generation of generalized descriptions for genotype combinations and assessment of cumulative genetic influence on a given trait. To optimize database construction, only genotype combinations identified in existing genome-wide SNP microarray datasets (1,415 genotyping profiles) were included rather than generating all theoretically possible combinations. As additional genomic data become available, the database is automatically expanded independently for each evaluated

characteristic.

Pretreatment of Genotyping Data

Individual genome-wide SNP microarray genotyping data was obtained using Illumina iScan platform and Infinium Global Screening Array Kit (654,027 SNPs) as described in Material and Methods section.

For quality control raw data were treated by GenomeStudio v.2010.3, and then exported and reformatted into PLINK-compatible files (.map and .ped). Subsequently, using PLINK v1.9, genotype information across 654,027 SNP loci per individual is parsed and saved into separate text-based datasets (Figure 3). Each record represents a single SNP and includes its rs identifier, chromosomal location, genomic coordinate, and corresponding genotype.

For the construction of the athlete genetic passport, only a predefined subset of 97 SNPs is required. Direct processing of full-scale datasets exceeding 600,000 entries per individual is therefore computationally inefficient, especially considering that a single PLINK dataset may include 24 to 96 samples simultaneously. To address this limitation, a dedicated preprocessing script is applied to extract only the target SNP panel, generating reduced files for each individual containing the selected 97 markers. This step significantly reduces computational load and accelerates downstream analytical proce-

```
# This data file generated by PLINK at: Fri Oct 20 17:03:01 2023
#
# Below is a text version of your data. Fields are TAB-separated.
# Each line corresponds to a single SNP. For each SNP, we provide its
# identifier, its location on a reference human genome, and the genotype call.
# For further information (e.g. which reference build was used), consult the
# original source of your data.
#
# rsid      chromosome    position    genotype
rs9651229    1          567667    CC
rs9701872    1          568208    TT
rs11497407   1          568527    GG
```

Figure 3. Example of a text file fragment containing SNP genotyping records

dures, including database integration and passport assembly.

To enable efficient storage and rapid retrieval of genotype information, the filtered data are further organized into a structured Excel-based matrix. In this format, rows correspond to rsIDs of the selected 97 SNP markers, while columns represent individual participants. Each cell stores the genotype value for a given SNP–individual pair. This tabular representation eliminates the need for repeated parsing of raw text files and facilitates fast access to genotype information.

The resulting matrix forms the core dataset for automated construction of SNP combination profiles and represents the initial stage of the athlete genetic passport pipeline. When new PLINK datasets are introduced, the matrix is dynamically extended by appending additional columns corresponding to newly genotyped individuals.

Extraction of SNP genotyping data

The initial stage of the workflow involves automated extraction of SNP text files from PLINK datasets. Because PLINK v1.9 operates through the command-line interface and is designed for simultaneous processing of large datasets, the Linux implementation was selected as the most effective environment. To automate extraction procedures, a Python script named `autopl原因ink.py` was developed. The script uses the `os` module [15] and the `Path` class from the `pathlib` library [18], which are particularly suitable for Linux-based file system operations and command-line execution. The `os` module allows direct execution of terminal commands from program code, including folder creation (`mkdir`) and automated execution of PLINK commands. The `Path` module is used to define and manage paths to source data files, particularly `.ped` files. The workflow consists of cyclic and non-cyclic components, with the cyclic component representing the central functional element of the script. During processing, copies of the original files are generated for each operation, ensuring preservation of the source data and eliminating the possibility of accidental information loss. The non-cyclic component prepares source datasets for downstream analysis. PLINK commands `-update-map` and `-update-name` are applied to correctly assign rsIDs to SNP markers according to the auxiliary Illumina annotation file (GSA-24v3-0_A1_b151_rsids.txt; Infinium Global Screening Array v3.0 (Illumina, USA) Loci Name to rsID Conversion File). After rsID correction, duplicate entries are identified and removed. Repeated rows are detected using the Linux command `cut`, whereas duplicate SNPs are excluded using the PLINK option `-exclude`. Following these procedures, extraction of sample-specific information becomes possible. The cyclic component performs automated extraction of genotype data for each sample individually. During execution, PLINK generates a `.fam` file containing identifiers of all samples present in the source dataset.

The program reads this file, stores the list of samples (typically 24–96 individuals), and sequentially extracts genotype information for each sample from the PLINK dataset. Extracted data are converted into a one-row-per-SNP format using the `--recode 23` options. The program then creates a dedicated folder for the sample and transfers both the final SNP file and all intermediate files generated during processing into that directory. As a result, all information related to a given participant is organized within a separate folder for simplified retrieval and storage.

Prior to development of the automated workflow, all procedures were performed manually. Updating rsIDs, removing duplicate entries, opening `.fam` files, copying sample identifiers, creating auxiliary text files, saving identifiers, and entering complex command-line instructions required substantial manual effort for each sample. Processing of a single sample previously required approximately 5–20 minutes, while one PLINK dataset could contain up to 96 individuals. The automated cyclic component now performs these procedures for all samples simultaneously. In addition to reducing processing time, the script minimizes errors associated with manual command entry, prevents data loss, and removes the need for repetitive re-entry of commands. Separate folders containing only SNP text files are also generated automatically for convenience, and filenames are assigned according to sample identifiers. Consequently, the developed script effectively eliminates the human factor from the extraction process, reducing the probability of errors to nearly zero and decreasing processing time for all samples within a PLINK dataset from several hours to approximately 1–2 minutes.

Selection of 97 SNP markers for the athlete genetic passport

Subsequent stages of data processing were implemented in R, using RStudio as the development environment. After launching the program, the user selects a folder containing previously extracted SNP text files derived from PLINK datasets. The program automatically detects the relevant files and initiates extraction of the target SNPs.

Initially, the software reads a predefined list of 97 target SNPs. It then extracts these SNPs from each participant-specific text file and generates a reduced file with the suffix “modified”. This procedure is repeated automatically for every sample included in the genotyping batch. The corresponding R code is presented in Figure 4.

As a result of filtering, the participant-specific files were reduced from approximately 16 MB to approximately 3 KB and contained only the genotype information required for passport generation. This reduction decreased the volume of data processed during subsequent stages, from approximately 654,027 SNP records to 97 target SNPs per participant.

```
all_rsid <- read.table('rsid_sport_all.txt', header = TRUE)
all_rsid <- unique(all_rsid)

for(i in c(1:length(sample_name))){
  rs_data <- read.table(paste0(dir, '/', ZapuskName(), '_rs_txt/', sample_name[i]), header = F)
  rs_data_modified <- dplyr::filter(rs_data, V1 %in% all_rsid$RSID)
  write.table(rs_data_modified, file = paste0('temp_mod_samples/', 'modified_', sample_name[i]), sep = "\t", row.names = F,
    col.names = c('rsid', 'chromosome', 'position', 'genotype'), quote = F)
}
```

Figure 4. R program workflow for extraction of target SNP markers

```

zapusk_all_samples <- read.table('all_genotypes.txt', header = T)
mod_sample_name <- list.files('temp_mod_samples/')
folder_name <- list.files(paste0(dir, '/samples'))

for (i in c(1:length(sample_name))){
  rs_table <- read.table(paste0('temp_mod_samples/', mod_sample_name[i]), header = T)
  zapusk_all_samples[folder_name[i]] <- rs_table$genotype
}
write.table(zapusk_all_samples, 'all_genotypes.txt', sep = '\t', row.names = F,
quote = F)

```

Figure 5. R script used for automated updating of the common genotype table

Updating the genetic variant table

Following formatting of SNP text files, genotype information is incorporated into the common variant table. For this purpose, the program reads the file together with the folder containing newly generated filtered SNP datasets. Each processed file is handled cyclically, allowing automatic addition of genotype information into the shared table (Figure 5). The syntax of the R programming language ensures that duplicate entries are not introduced, thereby maintaining uniqueness of all columns within the dataset.

Because row order in each text file exactly corresponds to the row structure of the general table, assignment errors involving mismatched SNP variants are effectively prevented. Consequently, the common genotype table is updated within less than one second using only new and unique information while preserving integrity of both rows and columns.

Removal of mirrored genotypes from the table

Certain SNPs may contain mirrored genotype representations such as GA/AG or TC/CT. Although biologically equivalent, these variants are interpreted by software as distinct values. To ensure accurate interpretation, the program identifies mirrored genotypes for selected rsIDs and converts them into standardized representations.

The corresponding code (Figure 6) evaluates rows associated with predefined rsIDs and reverses genotype order when necessary. For example, GA is converted to AG, TC to CT, and GT to TG. Because such inconsistencies occur only for a

```

genotypes <- read.table('all_genotypes.txt', header = T)
SNP <- read.table('mirrorgenotypesSNP/AG.txt', header = T)
for (i in c(1:nrow(SNP))){
  a <- which(genotypes$rsid == SNP[i,])
  genotypes[a,] <- replace(genotypes[a,], genotypes[a,] == "GA", "AG")
}
SNP <- read.table('mirrorgenotypesSNP/CT.txt', header = T)
for (i in c(1:nrow(SNP))){
  a <- which(genotypes$rsid == SNP[i,])
  genotypes[a,] <- replace(genotypes[a,], genotypes[a,] == "TC", "CT")
}
SNP <- read.table('mirrorgenotypesSNP/TG.txt', header = T)
for (i in c(1:nrow(SNP))){
  a <- which(genotypes$rsid == SNP[i,])
  genotypes[a,] <- replace(genotypes[a,], genotypes[a,] == "GT", "TG")
}

```

Figure 6. Program algorithm for standardization of mirrored genotype variants

limited number of SNPs, checking the entire dataset was considered computationally inefficient. Instead, only predefined loci known to contain mirrored variants are evaluated.

Configuration files

As noted previously, the software relies on a collection of external text-based configuration files that store fixed parameters required for correct workflow execution. Given the substantial number of such parameters, maintaining them separately from the source code improves flexibility and simplifies software maintenance. These configuration settings are loaded during the initialization stage before the program accesses databases containing genotype combinations and trait interpretation data (Figure 7).

The configuration files govern several key aspects of the system, including the arrangement of traits within information blocks, locations of databases and auxiliary resources, naming conventions for generated passport documents, and the processing order of evaluated athletic characteristics.

Identification of missing combinations of genetic variants and generation of new generalized interpretations

After genotype information from newly analyzed samples has been incorporated into the common variant table, the program evaluates whether previously unrecorded genotype combinations are present. This stage initiates assembly of the final athlete genetic passport. To generate the final report, the algorithm verifies whether predefined interpretations are available in the database for every genotype combination associated with each evaluated trait in all analyzed samples. Initially, the program defines general parameters including sample lists and the identifier of the current genotyping batch. These variables determine sample names and folder structures throughout the remaining stages of passport generation. The updated common genotype table is then loaded, and a dedicated directory corresponding to the genotyping batch is created for storage of the resulting documents. Following initialization, the software executes a multilevel algorithm organized through nested iteration structures that enable concurrent evaluation of multiple conditions. At the highest level, the main loop iterates over four distinct information databases in sequence. For each selected database, the system retrieves the corresponding trait set along with their predefined ordering as stored within

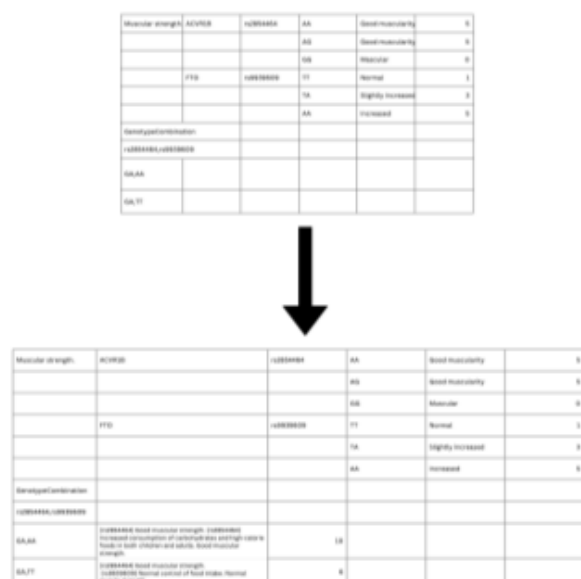
```

Orders <- scan(file = 'passportCreateAgruments/Orders.txt', what = '', sep = '\n')
Databases <- scan(file = 'passportCreateAgruments/Databases.txt', what = '', sep = '\n')
FeatureSheetDir <- scan(file = 'passportCreateAgruments/FeatureSheets.txt', what = '', sep = '\n')
Outputs <- scan(file = 'passportCreateAgruments/OutputFiles.txt', what = '', sep = '\n')
FeatureNames <- scan(file = 'passportCreateAgruments/FeatureNames.txt', what = '', sep = '\n')

```

Figure 7. Software module responsible for loading configuration settings

Similarly to the primary databases, the temporary database contains genotype-specific associations for each marker and three possible diploid genotypes within dedicated worksheets corresponding to individual traits. Information is therefore retrieved directly from the relevant sheets without requiring separate storage of all trait-related data. The algorithm writes interpretation descriptions for individual variants into adjacent empty fields corresponding to each genotype combination, thereby generating an integrated generalized interpretation based on cumulative effects of all variants.



The developed workflow integrates target-SNP extraction, genotype-matrix updating, genotype standardization, database matching, and generation of participant-specific reports. The complete sequence of automated data-processing stages is summarized in Figure 11.

```
for (s in c(1:4)){
  FinalList <- list()
  Feature <- scan(Orders[s], what = "", sep = '\n')
  FeatureNamesList <- scan(FeatureNames[s], what = '', sep = '\n')
  PassportShablon <- read.table("shablonPassport.txt", header = T)
  workProgress <- txtProgressBar(min = 0,max = length(Feature) * length(SamplesList), style = 3,
                                width = 25, char = '|||')
  a <- 0

  for (z in c(1:length(SamplesList))){

    for (i in c(1:length(Feature))){

      a <- a + 1
      Database <- read.xlsx(Databases[s],sheet = i)
      Featuresheet <- read.xlsx(paste0(FeaturesheetDir[s],Feature[i],'.xlsx'))
      FeaturesNPs <- filter(AllGenotypes, rsid %in% Featuresheet$rsid)

      SampleGenotype <- as.list(FeaturesNPs[SamplesList[z]])

      IndexConclution <- which(Database[[1]] %in% stri_paste(SampleGenotype[[1]], collapse = ','))
      if (length(IndexConclution) > 0){
        PassportShablon[1,2] <- Database[IndexConclution,2]
        PassportShablon[1,1] <- FeatureNamesList[i]
        PassportShablon[1,3] <- Database[IndexConclution,3]
        PassportShablon[1,4] <- Database[IndexConclution,4]
        PassportShablon[1,5] <- Database[str_which(Database[,1], 'GenotypeCombination') + 1,i]
      }

      setTxtProgressBar(workProgress, a)
    }

    FinalList <- append(FinalList, list(PassportShablon))
    names(FinalList)[z] <- SamplesList[z]
  }

  write.xlsx(FinalList, file = paste0("PreparedPassports/",txtname,"/",Outputs[s]))
}
```



Figure 10. Example layout of interpreted information in the final document

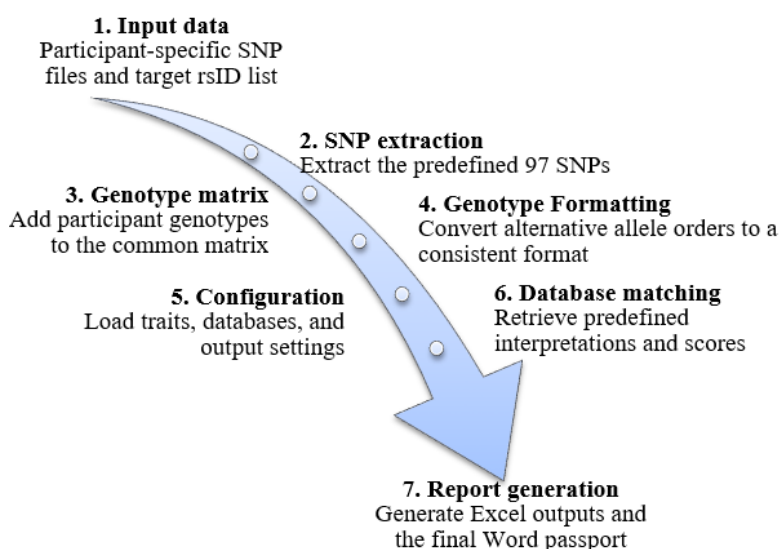


Figure 11. Complete automated workflow

As a result of this study, a software framework and several supporting scripts were developed using the RStudio environment and the Python programming language. The developed tools efficiently perform primary and secondary bioinformatic processing of genome-wide SNP genotyping data for generation of the athlete genetic passport. Databases containing relevant genetic information are created and updated automatically. Sorting and bioinformatic processing of genome-wide SNP genotyping data became substantially more time efficient. Reducing human involvement in repetitive processes of collecting and comparing tables minimized the probability of errors, while workflows associated with SNP genotyping data processing became significantly more automated and simplified.

Automation of information extraction from genotyping files, database updates, and assembly of final genetic passport considerably reduced data processing time. Complete genetic passport for an entire genotyping batch containing 24 to 96 samples can now be generated within 20 to 30 minutes. Previously, this process required from several days to one month.

ACKNOWLEDGMENTS

The authors thank the Institute of Genetics and Physiol-

ogy for providing access to the biobank resources used in this study. The authors also thank all participants whose samples and data contributed to this research.

AUTHOR CONTRIBUTIONS

Conceptualization, T.K. and S.M.; Methodology, Zh.Z.; Software, A.M.; Formal Analysis, T.K.; Investigation, Zh.Z. and A.M.; Resources, G.Zh.; Data Curation, N.A.; Writing – Original Draft Preparation, T.K., A.M. L.D; Writing – Review & Editing, T.K., L.D. and Zh.Z.; Visualization, A.M.; Supervision, L.D.; Project Administration, L.D.; Funding Acquisition, G.Zh., N.K. and S.M.

FUNDING

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under project BR28713313, «Development of an intelligent computational model for automating the solution of problems using supercomputer resources, with applications in bioinformatics and quantum computing and project BR18574139, «Development of complex system for training highly-qualified athletes and promising Olympic reserve for Kazakhstani priority sports using physiological genetic eval-

uation».

REFERENCES

- Bray M.S., Hagberg J.M., Pérusse L., et al. The human gene map for performance and health-related fitness phenotypes: The 2006–2007 update // *Medicine & Science in Sports & Exercise*. – 2009. – Vol. 41. – No. 1. – P. 35–73. – DOI: 10.1249/MSS.0b013e3181844179.
- Bouchard C., Hoffman E.P. (Eds.). *Genetic and Molecular Aspects of Sport Performance*. – London: Wiley, 2011.
- Vostrikova A., Pechenkina V., Danilova M., Boronnikova S., Kalendar R. Gene polymorphism and total genetic score in martial arts athletes with different athletic qualifications // *Genes*. – 2022. – Vol. 13. – No. 9. – Art. 1677. – DOI: 10.3390/genes13091677.
- Kazancı D., Polat T., Kaynar Ö., Bilici M.F., Tacal Aslan B., Ulucan K. PPARA and IL6: Exploring associations with athletic performance and genotype polymorphism // *Cellular and Molecular Biology*. – 2023. – Vol. 69. – No. 11. – P. 69–75. – DOI: 10.14715/cmb/2023.69.11.12.
- Leońska-Duniec A., Borczyk M., Piechota M., Korostyński M., Brodkiewicz A., Cieśczyk P. TTN variants are associated with physical performance and provide potential markers for sport-related phenotypes // *International Journal of Environmental Research and Public Health*. – 2022. – Vol. 19. – Art. 10173. – DOI: 10.3390/ijerph191610173.
- Petr M., Thiel D., Kateřina K., Brož P., Malý T., Zahálka F., et al. Speed and power-related gene polymorphisms associated with playing position in elite soccer players // *Biology of Sport*. – 2022. – Vol. 39. – No. 2. – P. 355–366. – DOI: 10.5114/biolsport.2022.105333.
- Ahmetov I.I., Hall E.C.R., Semenova E., Pranckevičienė E., Ginevičienė V. Advances in sports genomics // *Advances in Clinical Chemistry*. – 2022. – Vol. 107. – P. 215–263. – DOI: 10.1016/bs.acc.2021.07.004.
- Akhmetov I.I. *Molekulyarnaya genetika sporta (Molecular Genetics of Sport)*. – Moscow: Sovetskii Sport, 2009.
- Semenova E.A., Hall E.C.R., Ahmetov I.I. Genes and athletic performance: The 2023 update // *Genes*. – 2023. – Vol. 14. – No. 6. – Art. 1235. – DOI: 10.3390/genes14061235.
- Andryushchenko G.V. (Ed.). *Genetika i sport: Metodicheskie rekomendatsii (Genetics and Sport: Methodological Guidelines)*. – Moscow: Genotek, 2017.
- Thomson C.J., Hanna C.W., Carlson S.R., Rupert J.L. The –521 C/T variant in the dopamine-4-receptor gene (DRD4) is associated with skiing and snowboarding behavior // *Scandinavian Journal of Medicine & Science in Sports*. – 2013. – Vol. 23. – P. e108–e113. – DOI: 10.1111/sms.12031.
- Arnsten A.F., Scahill L., Findling R.L. Alpha2-adrenergic receptor agonists for the treatment of attention-deficit/hyperactivity disorder // *Journal of Child and Adolescent Psychopharmacology*. – 2007. – Vol. 17. – P. 393–406. – DOI: 10.1089/cap.2006.0098.
- Djansugurova L.B., Altynova N.K., Perfilieva A.V., Lebedeva L.P., Kumarbekov Zh.M., Kereyev T.N., Yergali K., Abylkassymova G.M., Sarsenova A.A., Khamdieva O.Kh., Skvortsova L.A., Mit N.V., Seisenbayeva A.S., Zhunussova G.S., Shadenova E.A., Bekmanov B.O. Method for Predicting Athletic Qualities Based on the Analysis of a Panel of 97 SNP Markers and Interpretation of Genetic Data Using a Point-Based Scoring of Characteristics. Utility Model Patent No. 9495. Bulletin No. 34. – 23 August 2024.
- Durinck S., Spellman P., Birney E., Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt // *Nature Protocols*. – 2009. – Vol. 4. – P. 1184–1191.
- Python Software Foundation. os - Miscellaneous operating system interfaces. Python 3.13.3 Documentation. – 2025. Available at: <https://docs.python.org/3/library/os.html>.
- Talbot J., Maves L. Skeletal muscle fiber type: using insights from muscle developmental biology to dissect targets for susceptibility and resistance to muscle disease // *Wiley Interdiscip. Rev. Dev. Biol.* – 2016. – Vol.5, Is. 4. – P. 518–34. doi: 10.1002/wdev.230.
- Mo M., Zhang Z., Wang X., Shen W., Zhang L., Lin S. Molecular mechanisms underlying the impact of muscle fiber types on meat quality in livestock and poultry // *Front. Vet. Sci.* – 2023. – Vol. 10. – P.1284551. DOI: 10.3389/fvets.2023.1284551
- Python Software Foundation. pathlib — Object-oriented filesystem paths. Python 3.13.3 Documentation. – 2025. Available at: <https://docs.python.org/3/library/pathlib.html>.

УДК 575-004.9

ОПТИМИЗАЦИЯ ОБРАБОТКИ ДАННЫХ ШИРОКОГЕНОМНОГО SNP-МИКРОЧИПОВОГО ГЕНОТИПИРОВАНИЯ ДЛЯ ФОРМИРОВАНИЯ ГЕНЕТИЧЕСКОГО ПРОФИЛЯ СПОРТСМЕНА

Т.Н. Кереев^{1,2}, А.В. Мустафаева^{1,2*}, Ж.А. Жаниязов^{1,2,3}, Н.К. Алтынова¹, Г.С. Жунусова¹, Л.Б. Джансугурова^{1,2*}, Н.М. Касымбек, С.Т. Мухамбетжанов

¹ РГП «Института генетики и физиологии» КН МНВО РК, Алматы, Казахстан

² Казахский Национальный Университет имени Аль-Фараби, Алматы, Казахстан

³ Казахский Национальный Исследовательский Технический Университет имени Сатпаева

*e-mail: mustafayeva_a@live.kaznu.kz (А.В.); leylad@mail.ru (Л.Б.)

АННОТАЦИЯ

На основании анализа международного опыта и результатов собственных исследований, посвящённых изучению связи между спортивными характеристиками человека и его генетическими особенностями по 97 SNP-маркерам, относящимся к 75 генам-кандидатам, была сформирована система интерпретации генотипов для генетического паспорта спортсмена. Для автоматизации обработки данных разработан программный комплекс, включающий набор скриптов на языках Python и R, реализованных в среде RStudio. Созданные инструменты обеспечивают автоматическое формирование и актуализацию баз данных генетической информации, что позволяет существенно сократить время подготовки генетического паспорта и минимизировать вероятность ошибок при обработке данных. Использование разработанных программных решений значительно повысило эффективность хранения, структурирования и анализа результатов широкогеномного SNP-микрочипового генотипирования. Автоматизация рутинных операций по сопоставлению и обработке больших массивов данных позволила практически исключить влияние человеческого фактора и упростить работу с генотипической информацией. Внедрение автоматизированных алгоритмов для первичной и вторичной обработки SNP-данных, интерпретации генотипов, обновления информационных баз и формирования итоговых генетических паспортов сократило продолжительность полного цикла обработки данных с нескольких месяцев до 20-30 минут.

Ключевые слова: спортивная генетика, генетический паспорт, широкогеномное SNP-микрочиповое генотипирование, однонуклеотидный полиморфизм, программный код.

ЭОЖ 575-004.9

СПОРТШЫНЫҢ ГЕНЕТИКАЛЫҚ ПРОФИЛІН ҚАЛЫПТАСТЫРУ ҮШІН ГЕНОМДЫҚ АУҚЫМДАҒЫ SNP-МИКРОЧИП ДЕРЕКТЕРІН ӨНДЕУДІ ОҢТАЙЛАНДЫРУ

Т.Н. Кереев^{1,2}, А.В. Мустафаева^{1,2*}, Ж.А. Жаниязов^{1,2,3}, Н.К. Алтынова¹, Г.С. Жунусова¹, Л.Б. Джансугурова^{1,2*}, Н.М. Қасымбек, С.Т. Мухамбетжанов

¹ ҚР ҒЖБМ ҒК «Генетика және физиология институты» РМК, Алматы, Қазақстан

² Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан

³ Қ. И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан

*e-mail: mustafayeva_a@live.kaznu.kz (А.В.); leylad@mail.ru (Л.Б.)

АННОТАЦИЯ

Әлемдік тәжірибе мен 97 SNP-полиморфизм (75 негізгі ген) бойынша спорттық мүмкіндіктер мен жеке генотип арасындағы өзара байланысты анықтауға арналған жеке зерттеулерді талдау негізінде спортшының генетикалық паспорты үшін генотиптердің ықтимал нұсқаларының интерпретациялары әзірленді. RStudio әзірлеу ортасы мен Python бағдарламалау тілі негізінде бағдарлама және бірнеше қосымша скрипттер жасалды. Құрастырылған бағдарламалар қажетті генетикалық акпаратпен деректер базаларын тиімді түрде құруға және жаңартуға мүмкіндік береді, бұл генетикалық паспортты әзірлеуге кететін уақытты қысқартып, қателік қаупін азайтады. Осы бағдарламалардың көмегімен толықгеномды микрочиптік генотиптеудің бастапқы деректерін сақтау, сұрыптау және өңдеу уақыт жағынан әлдеқайда тиімді болды. Кестелерді жинау және салыстырудың бірсарынды үдерісіндегі адами фактордың төмендеуі қателік қаупін іс жүзінде нөлге дейін азайтып, генотиптеу деректерімен жұмыс істеу үдерісін жеңілдетті. Геномдық ауқымдағы SNP-микрочип деректерін бастапқы және екінші реттік өңдеу, генотиптерді интерпретациялау, деректер базаларын жаңарту және дайын генетикалық паспорттарды жинақтау процестерін автоматтандыру деректерді өңдеу уақытын бірнеше айдан 20-30 минутқа дейін қысқартты.

Кілт сөздер: спорт генетикасы, спорттық генетика, генетикалық паспорт, геномдық ауқымдағы SNP-микрочип, бір нуклеотидті полиморфизм, бағдарламалық код.