UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
INSTITUTE OF MOLECULAR AND CELL BIOLOGY
DEPARTMENT OF EVOLUTIONARY BIOLOGY

Lidiia Zhytnik

# Detection of Natural Selection in Modern Human Populations on the basis of SNP Genotyping Data

BSc thesis

Supervisors: Georgi Hudjašov, MSc
Ene Metspalu, PhD

TARTU 2012

# TABLE OF CONTENTS

# ABBREVIATIONS

BMP – Bone Morphogenetic Protein

CEPH – Centre d'Etude du Polymorphisme Humain (Human Polymorphism Study Center)

CEU – Utah residents with Northern and Western European ancestry from the HapMap project

CHB+JPT – Japanese in Tokyo and Han Chinese in Beijing populations from HapMap project

CMS – Cronic Mountain Sickness

EHH – Extended Haplotype Homozygosity

HLA – Human Leukocyte Antigen

iHH – integrated Extended Haplotype Homozygosity

iHS – integrated Haplotype Score test

LD – Linkage Disequilibrium

LRH – Long-range Haplotype test

REHH – Relative Extended Haplotype Homozygosity

SD – Standard Deviation

STR – Short Tandem Repeats

UV(R) – Ultraviolet (radiation)

WGLRH – Whole-Genome Long Range Haplotype test

XP-EHH – Cross Population Extended Haplotype Homozygosity test

YRI – Yoruban in Ibandan (Nigeria) population from HapMap project

# INTRODUCTION

As well as all living creatures on the Earth, modern human species was shaped under the influence of natural selection. First anatomically modern humans originated in African continent approximately 200,000 years ago. After the Out of Africa migration modern *Homo sapiens* has settled different regions of Eurasia and gave rise to local European, Asian and Oceanian populations, and later to Americans. Adaptations to new environmental conditions, e.g. climate, UV radiation and diet, were one of the most important processes in the recent evolution of our species. Local evolutionary innovations have shaped different physiological features, like resistance to infectious diseases, skin pigmentation, lactose tolerance, adaptation to high-altitude, etc.

Modern technologies have provided revolutionary opportunities in the whole-genome sequencing and SNP genotyping. New genomic data induced the appearance of innovative methods for studying of human evolution and natural selection in our genome. Novel genome-wide selection scans are more robust, accurate, time-effective and sensitive, allowing to detect even very recent selection events in comparison with classical approaches.

Comprehension of the natural selection process action upon human populations not only reveals the evolutionary prehistory of our species, but also has very high importance for deeper understanding of current and potential diseases. Due to relatively low autosomal mutation rate our genes did not always have enough time to respond to the significant environmental changes which took place during the last few centuries. As a consequence modern humans tend to suffer from so-called “evolutionary diseases”: diabetes, obesity, cardiovascular diseases, etc.

The main aim of this study is to give a brief overview about new genome-wide selection scans, so-called haplotype-based methods. These methods are further applied to the test for the presence of recent positive selection in skin pigmentation associated *SLC24A5* gene using whole-genome genotyping data from world-wide human population sample.

# 1. LITERATURE OVERVIEW

## 1.1. Haplotype structure of human genome

### 1.1.1. Haplotypes

SNPs or single nucleotide polymorphisms are DNA sites, where two homologous chromosomes or members of single species differ from each other by a single nucleotide (Figure 1). Different variants of nucleotide substitution represent different forms of allele. It could be either bi- or multiallelic depending upon the number of detected nucleotide variants in the particular genomic position. Variants could be further classified to one of the following classes based on their frequency in population: low frequency variants (less than 0.5% frequency), rare variants (0.5 to 5% frequency) and common variants (more than 5% frequency in population) (The 1000 Genomes Project Consortium 2010). SNPs can be found in nearly all regions of human genome, including coding sequences of genes (exons), non-coding areas of genes (introns) and intergenic regions.

About 10 million SNPs are identified with common efforts of Human Genome Project, SNP Consortium and HapMap Project thus far. It is supposed that the number of nucleotide base pairs in the human genome is equal to 2.85 billion (International Human Genome Sequencing Consortium 2004) and, although the SNP density is not uniform across the genome, one SNP takes place every 300-1000 bp on average (The International HapMap Consortium 2003). The most variable regions in human genome are human leukocyte antigen (HLA) and sub-telomeric regions, while some gene-dense regions have significant decrease in the variability (The 1000 Genomes Project Consortium 2010).

Term haplotype comes from the Greek word *haploûs* meaning “single” or “simple” and describes a set of alleles on a single chromosome, which are transmitted together. Allelic states are determined by all kinds of DNA polymorphisms, including SNPs, short tandem repeats (STR), etc. Thus, haplotype could be defined as a specific combination of different alleles on one single chromosome, which passes from generation to generation unchanged (Figure 1). This ancestral association of alleles could be disrupted either by recombination or mutation event. Both these processes give rise to the new combination of allelic states – derived haplotype (Jobling et al. 2004).

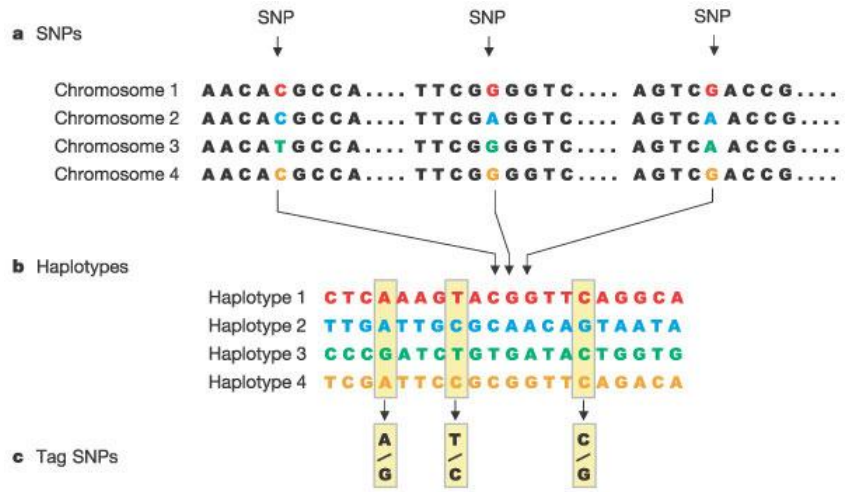


Figure 1. SNPs, haplotypes and tag SNPs (The International HapMap Consortium 2003).

a) SNPs. Variants of the same genomic region in different people. Most parts of the sequence are identical, however single nucleotide polymorphisms are found in three sites.

b) Haplotypes. Haplotype is specific combination of alleles, which is passed from generation to generation as a single unit until new haplotype arises due to the recombination or mutation. Depicted haplotypes include 20 SNPs, which were genotyped from 6 kbp long DNA sequences.

c) Tag SNPs. To identify the allelic states of all variable positions in the given set of haplotypes it is necessary to genotype only 3 tag SNPs.

### 1.1.2. LD (Linkage Disequilibrium)

Alleles, which are located at the nearby loci associate non-randomly and are inherited together or linked belonging to one common haplotype. This phenomenon is known as linkage disequilibrium (LD) (Ardlie et al. 2002). The association of linked alleles breaks down during recombination. The result of this process is a derived mosaic combination of markers (haplotype) and new unique chromosome (Figure 2). Rate of the linkage disequilibrium decay depends on the amount of historic recombination, which took place on the chromosome. Thus, LD could be useful in identifying of the age of the allele (Jobling et al. 2004).

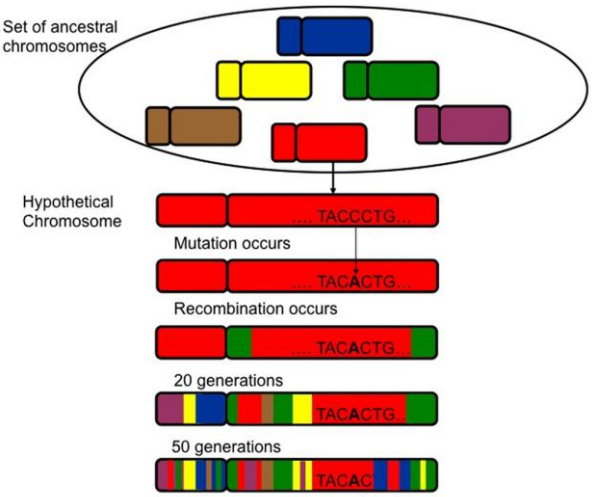


Figure 2. Linkage disequilibrium breakdown by recombination process. Different ancestral chromosomes in the population are shown in different colors. C to A mutation in hypothetical ancestral chromosome (red) gives rise to the new haplotype. Additional new derived chromosomes arise further during the recombination process (Dick et al. 2011).

This basically means that at the ends of the chr LD is a low number and will stick together better thus causing more variation, and in the middle of the chr LD is a high number and low variation

On the basis of recombination pattern across the genome, we can conclude that LD is usually low at telomere and high at centromere regions. Moreover LD is generally lower in genes connected to immune processes and neurophysiology. Genes which regulate cell cycle, RNA and DNA metabolism or DNA reparation system usually locate in the regions of strong LD and therefore stay conserved (The International HapMap Consortium 2005).

Lower LD equals more random variation and more SNPs at a given area; higher or strong LD equals less variation at that site

The standard measure of LD is D, which is the difference between an observed frequency of haplotype and the frequency of haplotype expected if the alleles are segregating at random (Lewontin 1964). D depends on allele frequencies and cannot be used if allele frequencies are equal to 0 or 1. Therefore, measures D′ and $r^2$ are in use. D′ is an absolute value of D divided by maximum theoretical value of observed allele frequencies. If D′=1 complete LD exists. D′ < 1 means ancestral LD is broken. $r^2$ or $\Delta^2$ is a correlation coefficient showing correlation of alleles in two loci. $r^2$ equals to 1, if allele frequencies stay the same and markers were not separated by recombination (Ardlie et al. 2002; Jobling et al. 2004).

### 1.1.3. Haplotype diversity

The haplotypes in the human genome have been shaped by molecular mechanisms of sexual reproduction and mutational processes. Different factors, including ionizing and ultraviolet radiation and chemical mutagens can cause genetic mutations. In addition, many biological processes play a significant role in mutagenesis: viral infections, errors in meiosis and DNA

replication, transposons, etc. All these factors break down the integrity of the genome and introduce new variation patterns. The most drastically mutations affect the non-recombining haploid portions of our genome – mtDNA and the majority of Y chromosome (Jobling et al. 2004). In the diploid portion of genome haplotype diversity also depends upon the processes of recombination and gene conversion. During the process of recombination ancestral DNA molecule is broken down and ligated with new homologous or non-homologous DNA molecule. Recombination process is distributed unequally in eukaryotes (Petes 2001). There are 25,000-50,000 areas with high probability of recombination or recombination hotspots (Myers et al. 2005). Hotspots are usually about 2 kb long and lie at the ends of the chromosome (The International HapMap Consortium 2005). Moreover certain correlation between recombination hotspots and GC-rich regions of the genome exists (Petes 2001).

Key principle to the pop-gen stats

On the population-level haplotype diversity is influenced by different evolutionary mechanisms, including natural selection and genetic drift. Another important evolutionary process is called genetic hitchhiking: when frequency of some advantageous SNP in the particular haplotype is being shaped by natural selection, e.g. SNP is being fixed by positive selection, then the frequency of all other linked SNPs in this haplotype will be also affected, and the whole haplotype will become fixed in the population. Thus, the overall haplotype diversity in given population will decrease, and the phenomenon is called a selective sweep (Barton 2000; Jobling et al. 2004; Smith and Haigh 2007).

Less variation in the population = positive selection

In addition, haplotype diversity is affected by different social and demographic factors: bottlenecks and population expansions, founder effects, mate choice, isolation or admixture of populations. For example, bottleneck could reduce the variety of haplotypes in the population, while admixture of populations could increase population's haplotype diversity, as two populations will enrich each other with new haplotype variants. Furthermore, the length and the diversity of haplotypes changes with time, while more and more recombination events occur in the course of evolution, the length of ancestral haplotype becomes smaller and the overall number of haplotypes in population increases. Thus, the larger the haplotype diversity the older the population is. The comparison of African, European and Asian samples detected, that 90% of all population-specific haplotypes is found in African continent. This, along with the largest haplotype diversity in Africa and very high similarity of Asian and European haplotypes, supports the hypothesis of the African origin of anatomically modern humans and points, that Africans is the oldest of three populations studied (Gabriel et al. 2002; Wang et al. 2002; Tishkoff and Verrelli 2003; Jobling et al. 2004).

Main Point

### 1.1.4. The structure of haplotype blocks

The overall structure of human genome is characterized by the presence of haplotype blocks or haploblocks, which are poor of recombination. These are large genomic regions where only few common haplotypes exist bordered by the points of recombination or recombination "hotspots". The haplotype block must have less than 5% of SNPs showing the evidence of historical recombination (upper bound of confidence interval of D' > 0.98 and lower bound > 0.7). These five percent are not taken into consideration, since they could arise through other biological and artificial factors, for example mutations or genotyping mistakes which also interfere haplotype pattern (Gabriel et al. 2002). Lack of recombination within blocks brings high linkage disequilibrium, so that haplotypes is transmitted in an undisrupted manner. However, recombination takes place between blocks, which makes interblock LD values small (Goldstein and Weale 2001).

Depending upon the population studied, half of the human genome exists in the blocks larger than 22 to 44 kb. Only three to five different haplotypes in each population describe around 90% of all genetic variation within each block (Gabriel et al. 2002). The fact that only very small number of different haplotypes describes the majority of genetic variation in the relatively large continuous sequence block is used to select tag SNP markers (Figure 1). Tag SNPs are in the linkage disequilibrium with other polymorphic positions in the particular haplotype block. Thus, typing just few carefully selected tag SNPs allows to detect the majority of variation within the particular haplotype block (The International HapMap Consortium 2003).

Principle behind Imputing SNPs

Sizes of haplotype blocks vary from 1 kb to 137 kb in different populations. Haploblocks of Africans are generally shorter (up to 96 kb) in comparison with European and Asian populations (up to 137 kb) emphasizing the „African origin" of modern human populations: the shorter haploblock is, the more recombinations took place and the longer time haploblock exists. Although many haploblocks are small in size, the majority of human genome consists of bigger blocks as illustrated on the Figure 3 (Gabriel et al. 2002).

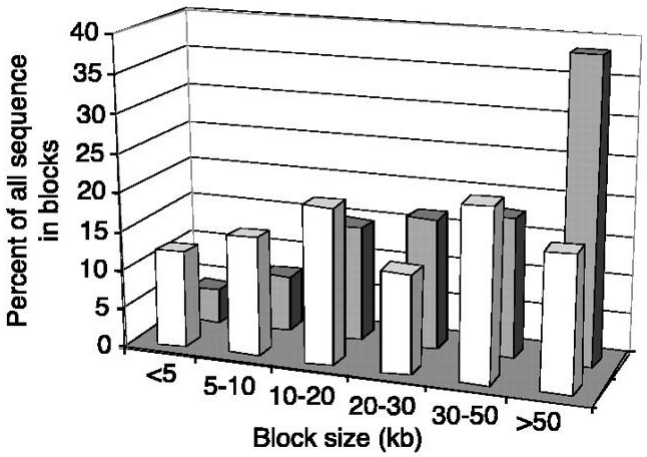


Figure 3. Distribution of genomic sequence in the blocks of different size. Grey bars correspond to European and Asian populations and white bars correspond to Yoruban and African American populations (Gabriel et al. 2002).

### 1.1.5. HapMap Project

From October of 2002 Haplotype Map of the human genome is supervised by International Organisation – the HapMap Project. The third version of HapMap Project includes information about ~1.5 million verified positions in the human genome from 1397 samples from 11 populations world-wide. Main aim of the project is to describe the variety of human genome. Our genomes are identical in 99.9% and studying of the rest 0.1% will reveal secrets of the influence of the environmental factors, drugs and diseases on the human organism (The International HapMap Consortium 2003). Initially, only three populations were used: Yoruba (Nigeria), China and Japan, and US citizens with East and North European origin. These populations represent three human races: Africans (YRI), Asians (CHB+JPT) and Europeans (CEU). Ten centers from different parts of the world proceeded genotyping of the SNPs. Tag SNPs were chosen and haplotypes were determined through standard measures of SNP association, D′ and $r^2$. All data of HapMap project is freely available.

## 1.2. Detecting natural selection on genome-wide level

As well as all other living organisms of the Earth, anatomically modern humans evolved under the influence of natural selection. There are different forms of natural selection acting upon the genome and different processes are leaving unique traces on the genetic variability of the species. Negative or purifying selection removes deleterious or harmful mutations from

Both positive and negative selection cause high LD? Maybe negative does high LD in individual and positive does high LD in population (within theses haploblocks)

the population, while, on the contrary, positive selection increases the probability of new advantageous genetic variant to be fixed (Jobling et al. 2004).

Depending on the data and hypothesis to be tested different selection tests could be used. Theses includes tests of polymorphisms within species (*Tajima's D, Fu and Li's D, $F_{st}$*), including haplotype based tests (*iHS, LRH, XP-EHH*), tests based on polymorphisms within species and the divergence between species (*Hudson–Kreitman–Aguade test, McDonald-Kreitman test*), interspecies tests (*dn/ds test*), etc. (Biswas and Akey 2006).

Different tests account for the different patterns of natural selection and should be used accordingly to the hypothesis studied. For example, *dn*/*ds* test is based on the ratio of rate of non-synonymous vs. synonymous substitution in pair of species (e.g. human-chimp pair) in a protein-coding gene. The *dn*/*ds* < 1 indicates the overabundance of synonymous substitutions and purifying selection acting upon protein-altering mutations in a gene, while > 1 indicates that some non-synonymous polymorphisms were fixed by positive selection (Nei and Gojobori 1986). As protein-altering mutations are very rare, the test is generally used for interspecies comparison, thus, to detect the natural selection in the evolutionary branch leading from one species to another and dating millions of years back, e.g. it could be used to detect selection in whole human lineage.

Another well-known test, Tajima's D, compares the number of mutation on the tips of evolutionary tree to the number of mutations in the external branches (Tajima 1989). This test is extremely affected by different demographic processes and should be used with great caution and well selected evolutionary model.

### 1.2.1. Tests based on length of the haplotypes

IMPORTANT!

Genetic variability is influenced not only by natural selection, but also by demographic processes. For example, great reduction in population size, bottleneck, will significantly decrease the genetic variability and thus randomly increase the frequency of some alleles. This will mimic the process of positive selection acting upon the same polymorphisms. One possibility to discriminate between two processes is to compare the studied genetic variability with other loci from the same population following the logics that natural selection will influence only certain regions, while complex demographic history will affect whole genome. The aim of current study is to give a brief overview on different haplotype based tests, also known as long-range haplotype tests. These tests were recently developed with the advance of whole-genome genotyping methods. They allow to exclude the influence of demographic

If it is positive selection, the phenomenon will only be found in that region; if it is complex demographic history (bottleneck) then you will see it everywhere in the genome.

processes, as background genetic variation from the same population is used as a reference. On the contrary to other tests, haplotype based statistics detect the traces of more recent natural selection (less than 30 kya or even 10 kya). They are looking for the long haplotypes with high frequency in population, which could be a result of selection for some advantageous allele. After the original fixation event the recombination process will slowly destroy extended haplotype block, thus the more time will pass since the time of fixation, the more likely the selection event will stay undetected (Sabeti et al. 2002; Jobling et al. 2004; Biswas and Akey 2006; Sabeti et al. 2006; Sabeti et al. 2007).

Haplotype based tests are flexible to choose genetic markers. All kinds of polymorphisms are accepted. In addition, they are generally regarded as more sensitive than other methods. These tests succeeded to identify natural selection in regions, where other methods could not find any footprints of selection from the same data. Disadvantages of the tests are their time depth, as long haplotypes tend to decay by recombination. For example, in 30,000 years, one crossing-over occurs approximately in every 100 kb of sequence (Sabeti et al. 2006).

The major haplotype based tests are *Long-Range Haplotype test* (LRH), *integrated Haplotype Score* (iHS) and *Cross Population Extended Haplotype Homozygosity* (XP-EHH).

Low LD = more variation and less selection
High LD = less variation and more selection (both positive and negative!)
NOTE: not 100% certain of this

### 1.2.1.1. LRH – Long-range haplotype test

Original long-range haplotype test was based on the measure of LD as calculated by EHH statistics, Extended Haplotype Homozygosity (Sabeti et al. 2002). First, core haplotype, which is a cluster (haplotype block) of SNPs with very low historical recombination, is identified. Then more distant SNPs upstream and downstream from core haplotype are added to the analysis one by one, this is referred as extended haplotype. For example, SNPs as far as 0.5-1 Mb upstream and downstream from the core region are usually included. EHH between each additional SNP and core haplotype is calculated as probability that two random chromosomes carrying core haplotype are identical by descent. Thus, EHH describes the transmission of extended haplotype without recombination or LD decay of the haplotype. If *t* is a studied haplotype, its EHH is calculated as follows:

1) Define core haplotype (use genetic map—low cM/Mb)
2) Count number with that haplotype
3) Find ext-hap for all indv
4) Get target ext-hap
5) Count indv with ext-hap of target
6) Count num of unique ext-hap

In a population, you have a core haplotype and individuals with that haplotype (c)
Total individuals in population (e)
Looking at an individual you see how many are the same as that one (s)

$$EHH_t = \frac{\sum_{i=1}^{s} \binom{e_{ti}}{2}}{\binom{c_t}{2}}$$

Where *c* is a number of samples with core haplotype, *e* is a number of samples with extended haplotype, *s* is a number of unique extended haplotypes.

* Denominator is the total number of combinations that can be made with a given phenotype
* For each of the unique extended haplotypes you see how many make a pair within set and set that over total number of pairwise combinations (I think)

If EHH is equal to 0 all haplotypes are different, EHH with value 1 shows that all haplotypes are identical (Sabeti *et al*., 2002). The higher EHH value is, the less recombination happened to the core haplotype and the younger core haplotype is. EHH statistics could be visualized as bifurcating diagram: tree is rooted in core haplotype, recombination events are displayed as splits, and tips of the tree correspond to unique extended haplotypes; branch thickness corresponds to the frequency of particular haplotype in the population (Figure 4a).

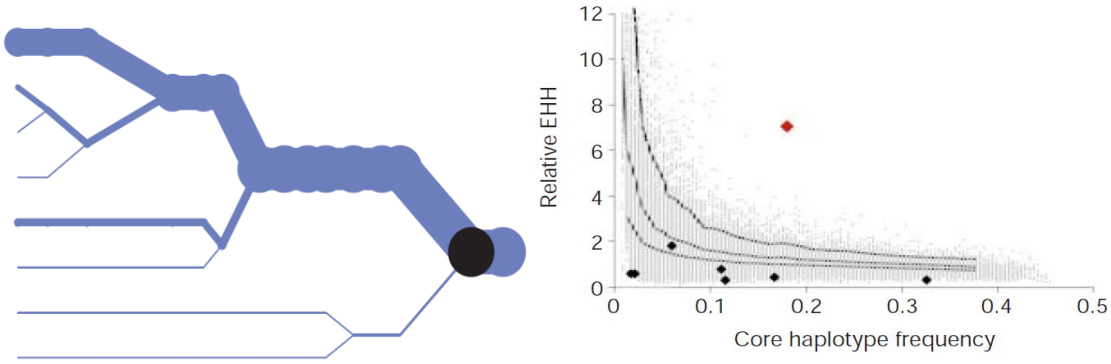


Figure 4. Extended Haplotype Homozygosity (EHH) diagram and outlying locus showing traces of positive selection in CH8 haplotype of *G6PD* gene (adapted from Sabeti et al. 2002).

a) Bifurcating diagram showing Extended Haplotype Homozygosity (EHH). Core haplotype is marked as black dot. Branches correspond to unique haplotypes and branch thickness corresponds to haplotype frequency in the population. Unusually long and frequent haplotype (uppermost branch) could point to the action of positive selection.

b) Relative EHH value plotted as a function of core haplotype frequency. Diamonds correspond to the empirical loci. *G6PD*-CH8 haplotype marked with red diamond has unusually high frequency and REHH value pointing to the past selection process. Area shaded in grey corresponds to the results of neutral *in silico* simulations with 95th, 75th and 50th percentiles are shown as black lines.

From what I can gather, REHH just compares one EHHt to all other EHHt; where you first add all possible EHHt together then compare a particular EHHt to the summation of all other EHHt

The formal test for selection uses relative EHH value (REHH), which is $EHH_t / \overline{EHH}$, where $\overline{EHH}$ refers to EHH from all other core haplotypes combined. High REHH values and high allele frequency could point to the action of positive selection on some particular core haplotype (Figure 4b). To explore how unusual particular REHH value is it could be also compared with *in silico* simulated neutral loci given the appropriate evolutionary model.

Further development of LRH is *Whole-Genome Long Range Haplotype test* (WGLRH). In this analysis empirical REHH values are compared to the distribution of whole-genome REHH values from alleles having similar frequency as studied one. Furthermore, alleles,

which undergone selection process, are likely to be new (derived) as opposed to ancestral alleles shared across hominine lineage. Thus, to decrease the false positive rate and improve computational speed it is appropriate to analyze only human derived alleles with high frequency (Zhang et al. 2006).

### 1.2.1.2. iHS – Integrated Haplotype Score test

Another tool for detection of natural selection is *Integrated Haplotype Score*, also known as iHS test (Voight et al. 2006). iHS was established on the basis of EHH statistics. This test allows for the identification of beneficial alleles, which was driven up by selection process up to intermediate frequency in population. The algorithm of the test is the following: original LRH test measures EHH statistics as function of distance from the core site, which could be plotted as a curve. iHS test integrates the area under such curve from the particular core site to the genomic position, where EHH reaches 0.05. This is done in both 5' and 3' directions, and for ancestral ($iHH_A$) and derived ($iHH_B$) alleles separately. The obtained statistics is:

$$unstandardized\ iHS = ln\left(\frac{iHH_A}{iHH_B}\right)$$

Ratio equal to 1 reflects the similar area under the curve for both ancestral and derived alleles (unstandardized iHS = 0) and points to the similar haplotype pattern on both chromosomes. Large negative or positive values points to the unusually long haplotypes carrying derived or ancestral allele, respectively. The unstandardized iHS is further compared to the genome-wide data from the same population to look for the outlying results, which could point to the action of positive selection. This allows to exclude the influence of different demographic processes and done by standardization procedure with alleles binned by their frequency:

$$iHS = \frac{ln\left(\frac{iHH_A}{iHH_B}\right) - E_p\left[ln\left(\frac{iHH_A}{iHH_B}\right)\right]}{SD_p\left[ln\left(\frac{iHH_A}{iHH_B}\right)\right]}$$

Where $E_p$ is the expected (average over the genome) iHS value of alleles with frequency $p$, and $SD_p$ is standard deviation of iHS values of alleles with frequency $p$.

Interesting plan of attack to the problem

It is more effective to look for the number of consecutive SNPs with high iHS score rather than analyze each SNP separately as the frequency of markers nearby to the selected site also increases due to the hitchhiking effect. This could be done by dividing the genome into non-overlapping windows (e.g. 200 kb in length) and calculating the ratio of SNPs with |iHS| > 2 for each window. Top 1% of distribution is used as a significance cutoff, thus only markers

belonging to the 99th percentile are generally considered as outliers showing traces of past selection (Voight et al. 2006; Pickrell et al. 2009).

The effectiveness of this test is supported by several lines of evidences. For example, the overabundance of extreme iHS score is found in those regions, which are also pinned up by various within species polymorphisms based tests and other haplotype length based tests and, as could be expected, also in genic regions (Voight et al. 2006; Sabeti et al. 2007).

### 1.2.1.3. XP-EHH – The Cross Population Extended Haplotype Homozygosity test

The next development of EHH is XP-EHH statistics or *Cross Population Extended Haplotype Homozygosity*. It detects the selection of alleles, which already reached the fixation in studied population (A), but remain polymorphic in reference population (B). Thus, it detects complete selective sweeps, while classical LRH and iHS is able to detect only partial fixation of selected alleles. The latter could be observed as a result of local adaptations to new environmental conditions (Sabeti et al. 2007). The algorithm of XP-EHH test is as follows: all SNPs 1 Mb upstream and downstream from the core SNP are considered and SNP X with EHH between 0.03 and 0.05 with respect to all chromosomes in both populations A and B is picked up. The lower EHH value is, the more recombination has occurred. Thus, picking up SNP X defines the most distal position from the core SNP, where no more long haplotypes exists. Next, EHH values for all SNPs between core regions and SNP X are calculated for population A and B separately. These values are further integrated as in iHS test, yielding $I_A$ and $I_B$ estimators. XP-EHH statistics defined as:

Is the EHH core defined by one

$$XP\text{-}EHH = ln\left(\frac{I_A}{I_B}\right)$$

XP-EHH is calculated in both directions, binned by allele frequency and normalized by whole-genome data. Analysis could be performed in non-overlapping windows (e.g. 200 kb in length); in this case maximum XP-EHH score in each window is used. Extremely positive or negative outlying XP-EHH scores show, that some particular haplotype is much longer in one population with respect to another, thus pointing to the positive selection for this particular haplotype. Positive score points to the selection in population A (studied) and negative in population B (reference). XP-EHH score could be calculated across the whole genome and top 99th or 99.5th percentiles could be used as significance cutoff values. African populations, e.g. Bantu, are usually used as a reference for non-African groups and European population is used as a reference for Africans (Sabeti et al. 2007; Pickrell et al. 2009).

Simulations show that XP-EHH has almost twice as much statistical power to detect complete allele fixation (selective sweep) in comparison to iHS test (0.97 vs. 0.48) and even more superior to LRH test (0.97 vs. 0.30, giving the emergence of selected allele 15 kya and European population demographic model) (Sabeti et al. 2007).

## 1.3. Examples of natural selection in human genome

### 1.3.1. *G6PD* gene

*G6PD* gene, which codes for glucose-6-phosphate dehydrogenase, is crucial in glucose metabolism, but persons with certain *G6PD* mutations have about 50% reduction in risk of malaria disease (Ruwende et al. 1995). Enzyme deficiency is common in many malarious countries, but not elsewhere. LRH test detected that A-202 allele of the *G6PD* gene increases resistance of the organism to *Plasmodium falciparum* (Sabeti et al. 2002). Research samples were taken from 3 African and 2 non-African populations. Core region of 15 kb was detected in the *G6PD* gene and 11 extended SNPs were genotyped. Nine haplotypes were identified (*G6PD*-CH1 to 9) with *G6PD*-CH8 haplotype carrying A-202 allele. This haplotype is common in African populations (18%), but not outside this continent. It has very high EHH value, which is 0.38 at 413 kb downstream from the core haplotype. Comparison to simulated neutral data also yielded highly significant differences for all African populations studied, proving the presence of natural selection in *G6PD* gene in Africa (Figure 4). The date of origin of resistant variant was estimated about 2500 BP. Furthermore, traditional tests (Tajima's D, *dn*/*ds*, etc.) were used on the same dataset and none of them detected any deviations from neutrality. This again underlines the weakness of classical tests to detect very recent selection events (Sabeti et al. 2002).

### 1.3.2. *LCT* gene

Lactase is involved in the hydrolysis of the disaccharide lactose into galactose and glucose monomers and is essential for digestion of milk. Certain alleles of *LCT* gene are connected to adult lactase persistence, the ability to digest milk in adulthood. Lactose intolerance frequency varies largely among different populations, ranging from up to 15% in northern Europeans to almost complete fixation in Southeast Asians and Native Americans. Population genetic studies suggest that lactase persistence has developed under effect of positive natural selection (Simoons 1970; Flatz 1987; Hollox et al. 2001; Poulter et al. 2003). Hypothesis was checked

with genetic analysis of *LCT* locus. Data was taken from representatives of European American, African American, East Asian and Scandinavian populations. According to LRH test results, haplotype with persistence-associated allele T-13910 (and A-22018) is found in 77% representatives of European American population and has highly unusual long-range LD extended for more than 800 kb giving the strong evidence for the presence of natural selection in *LCT* gene in Europe (Bersaglieri et al. 2004). *LCT* locus was also identified as one of the strongest candidates for recent positive selection in European population by XP-EHH and iHS analysis (Voight et al. 2006; Sabeti et al. 2007) Moreover, separate studies of African populations found three new independent alleles (C-14010, G-13915 and G-13907) associated with lactase persistence in this continent pointing to the convergent evolution of this trait in different human populations as the result from independent animal domestication and milk consumption in adulthood. Average homozygous tract length in lactase-persistent T/T-13910 homozygotes is 1.4 Mb in comparison to 1900 bp in C/C-13910 non-persistent Eurasian individuals. Lactase-persistent individuals from Africa, which carry beneficial C-14010 allele, have even longer span of LD, up to 2.9 Mb. This is the longest span of identity detected thus far in global HapMap data. The selection against African C-14010 allele is statistically significant in comparison to various simulated neutral models and also to genome-wide distribution of iHS scores from Yoruban HapMap data. Moreover, three African lactase-persistent alleles show 18 to 30% higher lactase expression profiles than non-persistent alleles (Tishkoff et al. 2007).

### 1.3.3. *SCA2* gene

Unusual patterns of strong LD were identified in *SCA2* gene. Representatives of Utah population with European ancestry showed extremely long haploblock on chromosome 12, 1.2 Mb long including 168 common polymorphisms with minor allele frequency more than 0.05 and average |D'| of 0.91. LRH test picked out a core region with seven tag SNPs inside or nearby *SCA2* gene and CH-1 haplotype was selected as a candidate of the past selection. This haplotype is very common in Europe, reaching 73% frequency. Although the molecular mechanisms of selection for such unusually long haplotype in not clear, CAG triplet expansions of spinocerebellar ataxia type 2 gene (*SCA2*) lead to severe neurodegenerative disorder. Gene itself is known to be involved in regulated cell death. Healthy controls have 14-31, while patients have more than 31 repeats and normal European CH-1 haplotype is made of 20 repeats (Yu et al. 2005).

### 1.3.4. *BMP3* gene

One of the substantial phenotypic differences between populations of modern humans lies within our skeletal system and include body mass, height, and craniofacial dimensions. *BMP3* is a human osteogenin gene, which takes part in the osteogenesis or bone formation. Osteogenin is an antagonist for osteogenic BMPs (bone morphogenetic proteins) and also negative determinant of bone mineral density, which plays crucial role in skeletal development. *BMP3* knockout mice have increased bone mass. Human *BMP3* gene is claimed to be a target of positive natural selection, although the molecular mechanism of this adaptation is not clear yet. Long-range haplotype test was used to confirm presence of selection in *BMP3* gene. Eleven tag SNPs were genotyped in core haplotypes of three populations (YRI, CEU, CHB+JPT) in order to reveal pattern of haplotype homozygosity in *BMP3* gene. One haplotype has very high frequency in European population, 79%, and its REHH and iHS values are unusually high (Wu et al. 2010).

### 1.3.5. Genes associated with adaptation to high-altitude

Tibetan populations, which inhabit high-altitude environment, share unique physiological traits not found elsewhere among humans, like decreased arterial oxygen content and hemoglobin concentration, increased resting ventilation, high infant survival rate etc. For example, Tibetans living at 4000 meters have approximately 10-20% lower hemoglobin concentration (1 g/dL less on average) than lowlanders, living at 2500 meters and do not suffer from CMS (Chronic Mountain Sickness). Individuals have normal oxygen metabolism despite permanent hypoxia due to the highland environment. Tibetans had to go through the adaptation and fine-tuning of the existing oxygen-transport system to extreme environmental conditions after the initial settlement of the region in order to protect organism from a fatal influence of a permanent low oxygen concentration such as hypoxic damage or CMS (Simonson et al. 2010 and references within).

Several recent whole genome selection scans have identified number of genes, which shows unique variation pattern in comparison to other populations. First, about 250 candidate genes associated with hypoxia response were selected. Second, complete selective sweeps were detected by XP-EHH test via comparison of Tibetans to low-land populations from China and Japan and/or partial sweeps were detected by iHS statistics. Several candidate genes do have traces of positive selection in studied population with *EPAS1* and *EGLN1* genes showing positive results in both tests across separate studies (Simonson et al. 2010; Wang et al. 2011;

Simonson et al. 2012). These genes play central role in the activation of hypoxia-inducible genes and homeostasis of hypoxia inducible factor in hypoxia and normoxia processes. Furthermore, regression analysis showed that each additional copy of beneficial selected haplotype of *EGLN1* gene decreases hemoglobin concentration by 1.7 g/dl on average (Simonson et al. 2010). Permanently lower hemoglobin level helps to avoid such CMS complication as blood hyperviscosity, the result of overproduction of hemoglobin in response to high-altitude environment in non-adapted individuals. Other selected genes were enriched in categories connected to blood vessels development, including placental vascular network (*VEGFA*, *ANGPT1*, and *ANG2*), embryos and female gonads. Therefore, unique adaptations in these genes in Tibetans could also explain higher infant survival rate and heavy-births in this high-land population (Wang et al. 2011).

### 1.3.6. Skin pigmentation and selection in *SLC24A5* gene

Skin color is assumed to be one of the most quickly evolving traits in anatomically modern humans. Even two populations, which are considered as genetically close to each other, have an impressive difference in levels of skin pigmentation. While the interpopulation diversity for the majority of SNPs composes 10-15% of the total variation, the skin pigmentation diversity between different populations could reach up to 88% (Relethford 2002).

Pigmentation diversity depends on qualitative and quantitative differences of melanin content. Melanin is a pigment responsible for the skin, hair and eye color. It is a complex of biopolymers produced in melanocytes, which are situated in basal layers of epidermis. Melanocytes contain special organelles, melanosomes, where melanin is synthesized. Melanosomes have tendency to gather around the nucleus of the cell and protect it from the damaging ultraviolet radiation (UVR). Melanogenesis is a continuous process and includes several steps: tyrosinase converts tyrosine into dopaquinone. In the presence of cysteine, dopaquinone forms pheomelanin, which has a red to yellow color. If cysteine is absent, brown to black pigment, photoprotective eumelanin is synthesized. There are also substantial differences between localization and formation of melanosomes among skin phenotypes. Lighter skin has smaller less-pigmented melanosomes rich in pheomelanin, which are grouped together. Dark skin has larger, more pigmented and unaggregated melanosomes, which mainly consist of eumelanin (Jablonski and Chaplin 2000; Jablonski 2004; Parra 2007).

In 1953 Italian geographer Renatto Beasutti created a map of distribution of human skin pigmentation between different populations. Strong correlation between skin color and

geographical latitude was detected. The closer the population is to the equator (Australia, Melanesia, South and Southeast Asia and sub-Saharan Africa) the darker skin color it has and vice versa. Analysis of global satellite data and skin phenotype at different latitudes confirmed that skin color depends on the annual UVR (280-400 nm) level. For example, level of damaging UV radiation is high at equator and sub-tropics, thus imposing additional constraints on the population living there. Harmful effect of UVR could appear in melanomas. Sunburns damage individual's sweat glands making skin much more susceptible to infections. Moreover, vitamin B folate photolysis could play important role in evolution of human skin color. This biomolecule is essential for DNA synthesis, reparation and methylation and is extremely susceptible to degradation by UV radiation. Additional photoprotection in high-UVR areas is provided by excess of eumelanin. Thus, natural selection has favored higher photoprotective eumelanin concentrations and darker skin color among population in (near-) equatorial areas (Jablonski and Chaplin 2000; Parra 2007; Jablonski and Chaplin 2010).

Another important property of ultraviolet B radiation is its ability to promote vitamin D synthesis in two inner layers of the skin: *stratum basale* and *stratum spinosum*. Vitamin D improves bone mineral density, plays active role in prevention of autoimmune diseases and participates in cell growth and differentiation. Individuals with darker skin and higher amount of eumelanin need up to 10 times higher UV radiation dosage to produce physiologically adequate amount of vitamin D in comparison to lightly-pigmented person. Thus, skin pigmentation in some particular population is an evolutionary compromise – it must be dark enough to provide protection from damaging UVR, but light enough to promote the production of sufficent vitamin D amount. Our African ancestors encountered high level of UVR, thus were darkly pigmented. After the Out of Africa migration proto-Eurasian populations colonized new geographical areas and adapted to lower UV radiation level. The new direction of natural selection was determined: lighter skin color (Jablonski and Chaplin 2000; Parra 2007; Jablonski and Chaplin 2010).

Pigmentation is a polygenic trait. There are around 170 mouse genes and their human homologues identified, which control mouse coat color (Montoliu et al. 2012). Human genes, which are known to be involved in the natural skin control variation, include *MC1R, ASIP, ADAM17, TYR, TYRP1, DCT, SLC24A5,* etc. (Sturm 2009). The majority of them were identified by whole genome association studies, but some also by whole genome selection scans. *MC1R* gene coding for G-protein coupled receptor is widely known as associated with red hair in Europeans (Harding et al. 2000). Another outlier is *SLC24A5* gene, which belongs to a family of potassium-dependent sodium/calcium exchangers residing in the melanosome

membrane. This is an ortholog of zebrafish *slc24a5* also known as "golden" gene (Lamason et al. 2005). Mutation in *slc24a5* cause decreased amount of melanin, "golden" phenotype is lighter than wild type. Human *SLC24A5* expression is 10-fold higher in skin and eye than in other tissue. *SLC24A5* rs1426654 SNP codes for alanine or threonine at amino acid 111 position. It is one of the top-ranked ancestry informative markers and has one of the highest allele differences between Africans and Europeans in HapMap database: 93-100% of Africans, Native Americans and Asians have G allele (coding for alanine), while 99-100% of Europeans have A allele (coding for threonine) (Figure 5). Moreover, Europeans have on average 55-fold decrease in heterozygosity on chromosome 15 around *SLC24A5* gene showing evidence for the strong selective sweep in proto-European population as confirmed by XP-EHH and iHS statistics (Sabeti et al. 2007; Pickrell et al. 2009). Variation in rs1426654 could explain up to 40% of European-African skin color difference. Surprisingly, sharing of ancestral allele between heavily pigmented Africans and light-skinned East Asians points to different molecular mechanisms and convergent evolution of light skin color in European and East Asian populations as confirmed by number of studies (Lamason et al. 2005; Norton et al. 2007; Pickrell et al. 2009).

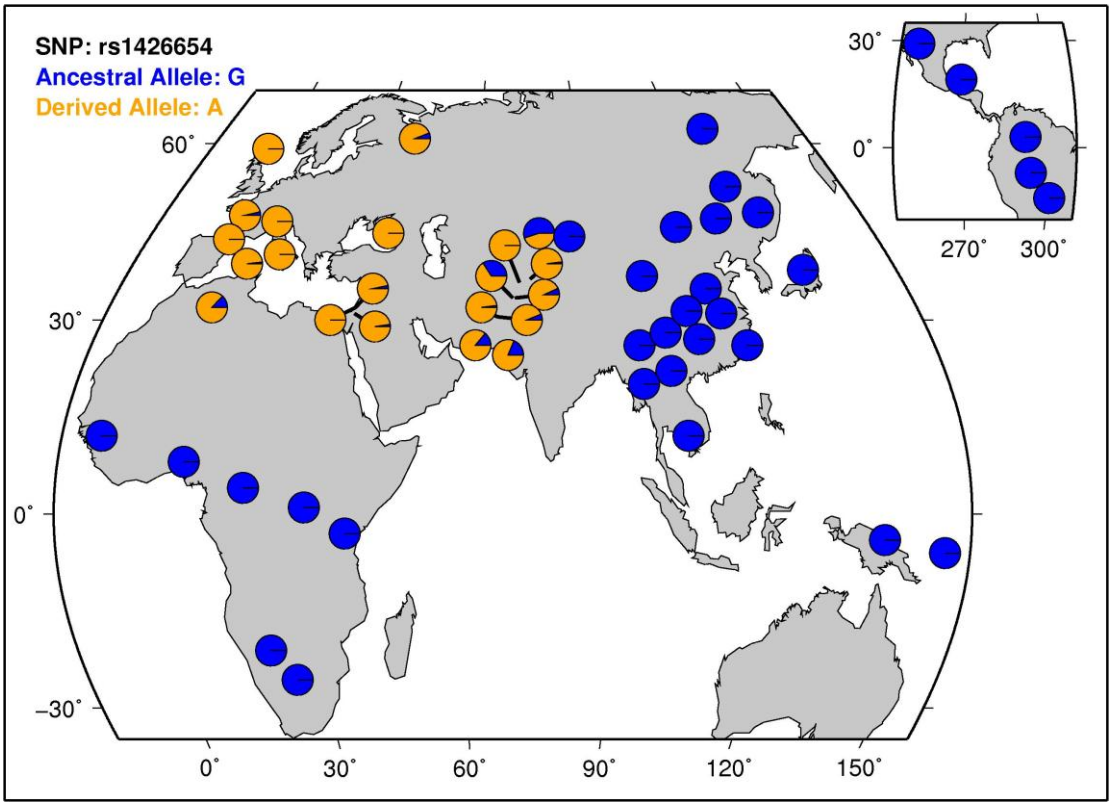


Figure 5. Frequency of *SLC24A5* rs1426654 alleles in different populations as visualized by HGDP selection browser (Coop et al. 2009; Pickrell et al. 2009; Pritchard et al. 2010). Ancestral and derived alleles are shown by blue and orange colors, respectively.

# 2. EXPERIMENTAL PART

## 2.1. Aim of the study

The main goal of current study is to confirm the hypothesis of presence of natural selection in human *SLC24A5* gene using the whole-genome genotyping data from world-wide population sample. Previous published data included long-range haplotype analysis of this gene in African (Yoruba), Asian (China and Japan) and European (Utah residents with Northern and Western European ancestry) populations (Lamason et al. 2005; Sabeti et al. 2007), and later whole-genome association study of UK residents with South Asian (India, Pakistan, Bangladesh, or Sri Lanka) ancestry (Stokowski et al. 2007). Significant iHS and XP-EHH results in this gene were also detected using CEPH sample from different world-wide populations (Pickrell et al. 2009), but, to our knowledge, no long-range haplotype analysis of *SLC24A5* gene was published on Indian population. Our updated sample set includes additional Indian, Oceanian and American samples. Current experimental work will test for the evidence of natural selection in *SLC24A5* gene using iHS and XP-EHH whole-genome selection scans.

## 2.2. Materials and methods

### 2.2.1. Genotyping data

Autosomal genotyping data from 348 samples (696 haplotypes) was used in current study. Based on the geographical affiliation samples were grouped into one of the following populations: Africa, Europe (including Near East), East Asia, India, Oceania and America (includes only Native American samples). Phased genotyping data of 337 samples was published previously (Li et al. 2008; Metspalu et al. 2011 and references within) and 11 new Indian samples genotyped using Illumina 610K bead array are reported here for the first time (see Supplementary Table 1 for additional sample details). Total number of samples in six geographic groups was as follows: 59 Africans, 33 Americans, 80 East Asians, 68 Europeans, 80 Indians (and Pakistan), and 28 Oceanians. Genotype to haplotype reconstruction (phasing) of all data was performed by Irene Gallego Romero (Department of Human Genetics, University of Chicago, USA).

### 2.2.2. Data processing

All data processing and computation was performed remotely on the University of Tartu aur.hpc.ut.ee computing cluster. The analysis algorithm and step-by-step workflow was as follows:

a. Complete chromosome 15 (15095 SNPs in total) including *SLC24A5* gene (genomic positions 48 413 169 to 48 434 869 as in GRCh37/hg19 human genome assembly) genotyping data was extracted from master-file in PLINK 1.07 software (Purcell et al. 2007; Purcell 2012) for each of the six populations separately;

b. PLINK output was further converted into *xpehh* and *ihs* software format using UNIX bash script coded by Irene Gallego Romero (Department of Human Genetics, University of Chicago, USA);

c. *xpehh* (coded by Joe Pickrell) and *ihs* (coded by Sridhar Kudaravalli) software were used to calculate XP-EHH and iHS statistics, respectively (Voight et al. 2006; Sabeti et al. 2007; Pickrell et al. 2009). Both packages were used with default settings and are available at http://hgdp.uchicago.edu/Software/. African Bantu population (N=19) was used as a reference for all non-African groups and Europeans was used as a reference for African group in XP-EHH analysis;

d. Raw iHS and XP-EHH values was binned by allele frequency and normalized against whole-genome allele data from similar derived allele frequency bin in Microsoft Excel 2010. Whole-genome data was calculated previously by Georgi Hudjašov (University of Tartu and Estonian Biocentre);

e. Normalized data was divided into 200-kb non-overlapping genomic windows and window containing *SLC25A5* gene was analyzed. Maximum normalized value of XP-EHH and proportion of |iHS > 2| in 200-kb window were used as test statistics;

f. Obtained XP-EHH and iHS scores were further compared with empirical whole-genome XP-EHH and iHS estimates from the same six sample sets (calculated previously by Georgi Hudjašov). Reference whole-genome data was also divided into 200-kb windows and binned by number of SNPs in each window (step size=20 SNPs); maximum value of XP-EHH and proportion of |iHS > 2| in each window were used. Genomic window containing *SLC24A5* gene was compared vs. whole-genome windows with similar number of SNPs (21 to 40). Percent rank of *SLC24A5* window against whole-genome distribution was calculated and 99th percentile was used as a significance cutoff.

## 2.3. Results and discussion

200-kb genomic window containing *SLC24A5* gene included 28 SNPs in total and spanned nucleotide positions 48 400 001 to 48 600 000 on chromosome 15 (GRCh37/hg19 human genome assembly).

Percent rank results of XP-EHH statistics for *SLC24A5* gene containing window are shown in Table 1. XP-EHH data points to the strong positive selection signal in European and Indian populations. This is also confirmed by highly negative XP-EHH score in Africa: XP-EHH test is directional, positive value points to the selection in studied and negative in reference population (Sabeti et al. 2007), and European population was used as a reference for Africans. Thus, bearing in mind that Europeans have strong positive selection signal, negative signal in Africa is also expected. Furthermore, 200-kb genomic window containing *SLC24A5* gene has the highest XP-EHH score in European population in our dataset (5.903) confirming previous findings (Smith et al. 2004; Lamason et al. 2005).

Table 1. Results of XP-EHH statistics. XP-EHH scores (maximum XP-EHH value) from 200-kb genomic window containing *SLC25A5* gene are shown for each studied population separately. Also shown percent rank for each respective window in the distribution of whole-genome 200-kb XP-EHH windows from the same sample set. Only those whole-genome windows, which contain 20 to 40 SNPs are considered in the percent rank calculation. Significant results passing 99th percentile values are marked in bold.

| Population | Normalized XP-EHH score | Percent rank of a window |
|---|---|---|
| Africa | -2.976 | 0.000 |
| America | 0.482 | 0.324 |
| East Asia | 1.275 | 0.649 |
| Europe | 5.903 | **1.000** |
| India | 3.816 | **0.995** |
| Oceania | 1.618 | 0.812 |

Results of iHS selection scan are shown in Table 2. None of the windows analyzed passed the top 1% significance cutoff, although European population is in the 96.5th percentile of the distribution and Indians showing second highest signal.

Table 2. Results of iHS statistics. iHS scores (proportion of |iHS > 2|) from 200-kb genomic window containing *SLC25A5* gene are shown for each studied population separately. Also shown percent rank for each respective window in the distribution of whole-genome 200-kb iHS windows from the same sample set. Only those whole-genome windows, which contain 20 to 40 SNPs are considered in the percent rank calculation.

| Population | Normalized iHS score | Percent rank of a window |
|---|---|---|
| Africa | 0.000 | 0.000 |
| America | 0.042 | 0.680 |
| East Asia | 0.040 | 0.648 |
| Europe | 0.214 | 0.965 |
| India | 0.074 | 0.770 |
| Oceania | 0.000 | 0.000 |

Discrepancy between XP-EHH and iHS results could be largely explained by the power of two tests to detect complete selection sweeps. Ten out of 28 alleles in *SLC24A5* window in our European population have reached fixation in respect to other samples (Supplementary Table 2). Power of iHS statistics to detect complete selective sweeps is weak, ranging from 0.59 to 0.38 for European population model and depending upon the time since selection event. On the other hand, XP-EHH statistics has almost as twice as power in the same conditions, ranging from 1 to 0.53 (Tables S2 and S3 in Sabeti et al. 2007). Furthermore, Europeans have about 100 kb long uninterrupted stretch of complete homozygosity starting from rs2433354 (position 48 414 969) to rs9920281 (position 48 514 309) in *SLC24A5* window in our sample confirming strong signal of selection and complete sweep as suggested previously (Lamason et al. 2005; Sabeti et al. 2007; Pickrell et al. 2009).

Genome-wide selection scans are based on the genotyping data and naturally cannot always pinpoint the exact polymorphism which was selected for. Thus, the selected allele in our European and Indian samples could be either one of the genotyped polymorphisms, or any other mutation in this region, which is in complete LD with SNPs genotyped. In latter case, "parasitic" alleles became fixed due to the process of genetic hitchhiking (Smith and Haigh 2007). Lamason et al. (2005) suggested that non-synonymous rs1426654 mutation was

selected for in *SLC24A5* gene. Although, we do not have this genotype in our database, we did detect very strong signal in *SLC24A5* region, once again underlining the pitfall of genotype-based analysis to detect the exact culprit of selection. To further investigate evolution of *SLC24A5* gene we need to apply complete gene, 5' and 3' UTR resequencing approach in world-wide population sample. This will allow for the exact identification of selected allele. Furthermore, functional studies are needed to understand molecular mechanism behind *SLC24A5* variation and its association with natural skin color differences among modern humans.

The presence of signal in both Europeans and Indians could reflect that selection took place before these two populations diverged. Alternatively, our data also does not rule out the possibility for convergent evolution of *SLC24A5* gene in two groups. In other words, we cannot exclude that different mutations were selected among European and Indian populations. The absence of signal in lightly pigmented East Asians also concurs with published data and points to the convergent evolution of light skin color in Europe and Asia (Lamason et al. 2005; Norton et al. 2007; Stokowski et al. 2007; Pickrell et al. 2009).

Our Indian sample is a mixture of individuals with both light and dark skin phenotypes. Stokowski et al. (2007) performed genome-wide association study using lightly and heavily pigmented South Asian cohorts from India, Pakistan, Bangladesh and Sri Lanka and found that *SLC24A5* rs1426654 polymorphism may explain > 30% of the variance between two pigmentation samples sets. Thus, results from our Indian sample could be even underestimated due to the "dilution" of selection signal for the light pigmentation by alleles coming from dark individuals. This important information must be considered during future study design and sample selection.

# CONCLUSION

Natural selection process leaves unique footprints in the genome allowing for the identification of spatial and temporal point of selection. Regions of the genome, which were influenced by action of positive selection have features specific for both ancient and new haplotypes: high extended haplotype homozygosity (EHH) and high frequency in population. High EHH reflects that recombination did not have enough time to destroy long LD produced by fixation of selected allele and hitchhiking of linked polymorphisms, while high population frequency points to the fitness advantage of the evolutionary innovation.

*SLC24A5* gene is associated with natural skin color variation in modern human populations and stands out among other candidate genes for recent positive selection in humans. The A allele of rs1426654 SNP in this gene is virtually fixed in European populations. It was hypothesized that selection for A allele explains up to 40% of difference in skin pigmentation between European and African populations. Lightly pigmented eastern Asians have ancestral G allele in the same position, which points to the convergent evolution of depigmentation in Europe and East Asia. To test for this hypothesis we have conducted XP-EHH and iHS genome-wide selection scans.

Our data includes 28 SNPs in *SLC24A5* 200-kb genomic window from 696 haplotypes from six world-wide populations: Africans, Americans, East Asians, Europeans, Indians and Oceanians. After normalization of the XP-EHH test results, strong signal of recent positive selection was defined in European (5.903, PR 1.000) and Indian (3.816, PR 0.995) populations. Negative values of XP-EHH statistics in Africans (-2.976, PR 0.000) also emphasize selection in reference (European) population. Although iHS test did not succeed to identify evident presence of selection, iHS scores for Europeans (0.214, PR 0.965) and Indians (0.074, PR 0.770) were the highest among other tested populations. This discrepancy between two tests could be explained by inefficiency of iHS in detection of already fixed or nearly fixed alleles.

From our results we can conclude that:

1. There was strong positive selection in *SLC24A5* gene in Europe.
2. There are also traces of positive selection in India, but due to the heterogeneity of this population the selection signal could be underestimated in our sample. This must be considered in future study design.

3. Absence of selection signal in East Asian sample points to the convergent evolution of light pigmentation between Europeans and East Asians.

For further examination of evolution of *SLC24A5* gene exact SNPs, which were favored by positive natural selection, must be identified. This could be done by complete resequencing procedure. Additional functional studies may allow for the identification of molecular mechanisms behind the *SLC24A5* variation and its association with different skin color phenotypes in humans.

# Loodusliku valiku hindamine genotüpiseerimise andmete põhjal kaasaegsetes inimpopulatsioonides

Lidiia Zhytnik

## Resümee

*Homo sapiens*, nagu kõik teised liigid, on kujunenud loodusliku valiku toime all. Anatoomiliselt kaasaegne inimene tekkis Aafrikas, kust asustas seejärel kogu muu maailma: Euroopa, Aasia ja Okeaania ning hiljem Ameerika. Uued keskkonna tingimused soodustasid lokaalset populatsioonide evolutsiooni. Loodusliku valiku toimel tekkisid uued kohastumused, nt. laktoosi tolerantsus, resistentsus teatud infektsioonilistele haigustele, muutused naha pigmentatsioonis ning adaptatsioon madalale hapniku konstentratsioonidele Tiibeti piirkonnas.

Loodusliku valiku toimimise tagajärjel erineb valiku all oleva genoomi osa struktuur üldisest taustast, mis evolutsioneerub neutraalse mustri järgi. Neutraalsest mustrist kõrvalekaldumiste detekteerimist ja analüüsi kasutatakse erinevates loodusliku valiku hindamistestides. Nende hulgas on liigisisesed polümorfismide testid (*Tajima's D, LRH, iHS, XP-EHH*), liigisiseste polümorfismide ja liigivahelise divergentsuse testid (*Hudson–Kreitman–Aguade test, McDonald-Kreitman test*) ja liikidevahelised testid (*dn/ds test*).

Haplotüüp on kõrvuti paiknevate alleelide kogum. Ühte haplotüüpi kuuluvad alleelid on omavahel aheldatud ning pärinevad koos. Rekombinatsioonid ja/või mutatsioonid antud haplotüübi piires rikuvad selle haplotüübi terviklikkust ning annavad aluse uute haplotüüpide tekkele. Mida noorem on haplotüüp, seda vähem rekombinatsiooni protsesse on ajas toimunud ja seda suurem on haplotüübi pikkus, mida hinnatatakse laienenud haplotüübi homosügootsusega (EHH, Extended Haplotype Homozygosity). Teine oluline näitaja loodusliku valiku hindamisel on alleelide sagedus. Mida suurem on alleeli sagedus populatsioonis, seda vanem on haplotüüp, kuid positiivse loodusliku valiku all olev alleel saavutab fikseerimise palju kiiremini. Seega, kõrge sagedusega suure pikkusega haplotüübi esinemine peegeldab loodusliku valiku toimet antud geneetilises piirkonnas.

Nende loodusliku valiku tunnuste põhjal olid välja töötatud nn. haplotüübil põhinevad testid. LRH (Long Range Haplotype test), iHS (integrated Haplotype Score) ja XP-EHH (Cross Population Extended Haplotype Homozygosity) testid on palju tundlikumad kui klassikalised ning võimaldavad detekteerida hiljutist positiivset valikut. Nende alumine ajaskaala on

piiratud ca 30 000 aastaga BP, hiljem esialgne pikk haplotüüp kaob rekombinatsioonide tõttu, ning teda on raske eristada neutraalsest genoomi taustast.

LRH või pika haplotüübi test põhineb EHH väärtuste ja alleelide sageduste võrdlemisel. iHS, ehk integreeritud haplotüübi väärtus, põhineb integreeritud EHH-väärtusel ning on uute (derived) ja vanade (ancestral) haplotüüpide iHS väärtuste suhe. iHS test on mõeldud mitte-täielikult fikseeritud alleelide hindamiseks. XP-EHH on populatsioonivaheline laiendatud haplotüübi homosügootsuse test. XP-EHH test lubab detekteerida alleeli, mis on täielikult fikseeritud ühes populatsioonis, kuid on polümorfne teises.

Hiljutise positiivse loodusliku valiku mõju oli leitud nt. *LCT*, *SCA2*, *G6PD* ja *BMP3* geenides. iHS ja XP-EHH testide abil oli tuvastatud valiku toime ka *SLC24A5* geenis, mille rs1426654 polümorfism on seotud heledama naha pigmentatsiooniga Euroopa inimpopulatsioonis ning mille varieeruvus kirjeldab kuni 40% naha värvi erinevust eurooplaste ja aafriklaste vahel.

Käesoleva töö eesmärgiks oli kontrollida loodusliku valiku esinemist *SLC24A5* geenis iHS ja XP-EHH testide abil kogu genoomi genotüpiseerimise andmestiku põhjal. Kokku analüüsiti 696 haplotüüpi kuuest inimpopulatisoonist: Aafrika, Ida Aasia, Ameerika, India, Euroopa, Okeaania. XP-EHH ja iHS testide tulemused viitavad tugevale positiivsele valikule Euroopa (5.903, PR 1.000) ja India (3.816, PR 0.995) populatsioonides. Negatiivne XP-EHH väärtus Aafrikas (-2.976, PR 0.000) kinnitab valiku olemasolu Euroopa referentspopulatsioonis. Paraku iHS test ei tuvastanud selektsiooni signaali mitte ühelgil testitud populatsioonil, kuid eurooplaste puhul saadud tulemus (0.214) oli 97. pertsentiili sees. iHS ja XP-EHH testide tulemuste erinevust võiks seletada iHS testi ebaefektiivsusega fikseeritud alleelide detekteerimisel.

Antud informatsiooni põhjal võime järeldada tugeva positiivse loodusliku valiku toimet *SLC24A5* geenis Euroopa ning India populatsioonides. Selektsiooni signaali puudumine ida-aasialaste seas viitab konvergentsele heleda naha evolutsioonile Euroopas ja Ida Aasias.

# REFERENCES

A) Magazines

(2003). "The International HapMap Project." Nature **426**(6968): 789-796.

(2004). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-945.

(2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-1320.

(2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Ardlie, K. G., L. Kruglyak, et al. (2002). "Patterns of linkage disequilibrium in the human genome." Nat Rev Genet **3**(4): 299-309.

Barton, N. H. (2000). "Genetic hitchhiking." Philos Trans R Soc Lond B Biol Sci **355**(1403): 1553-1562.

Bersaglieri, T., P. C. Sabeti, et al. (2004). "Genetic signatures of strong recent positive selection at the lactase gene." Am J Hum Genet **74**(6): 1111-1120.

Biswas, S. and J. M. Akey (2006). "Genomic insights into positive selection." Trends in Genetics **22**(8): 437-446.

Coop, G., J. K. Pickrell, et al. (2009). "The role of geography in human adaptation." PLoS Genet **5**(6): e1000500.

Dick, D., B. Riley, et al. (2011). "Incorporating Genetics into Your Studies: A Guide for Social Scientists." Frontiers in Psychiatry **2**.

Flatz, G. (1987). "Genetics of lactose digestion in humans." Adv Hum Genet **16**: 1-77.

Gabriel, S. B., S. F. Schaffner, et al. (2002). "The structure of haplotype blocks in the human genome." Science **296**(5576): 2225-2229.

Goldstein, D. B. and M. E. Weale (2001). "Population genomics: linkage disequilibrium holds the key." Curr Biol **11**(14): R576-579.

Harding, R. M., E. Healy, et al. (2000). "Evidence for variable selective pressures at MC1R." Am J Hum Genet **66**(4): 1351-1361.

Hollox, E. J., M. Poulter, et al. (2001). "Lactase haplotype diversity in the Old World." Am J Hum Genet **68**(1): 160-172.

Jablonski, N. G. (2004). "THE EVOLUTION OF HUMAN SKIN AND SKIN COLOR." Annual Review of Anthropology **33**(1): 585-623.

Jablonski, N. G. and G. Chaplin (2000). "The evolution of human skin coloration." J Hum Evol **39**(1): 57-106.

Jablonski, N. G. and G. Chaplin (2010). "Human skin pigmentation as an adaptation to UV radiation." Proceedings of the National Academy of Sciences **107**(Supplement 2): 8962-8968.

Lamason, R. L., M. A. Mohideen, et al. (2005). "SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans." Science **310**(5755): 1782-1786.

Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models." Genetics **49**(1): 49-67.

Montoliu, L., W. S. Oetting, et al. (2012). "Color Genes." from http://www.espcr.org/micemut/.

Myers, S., L. Bottolo, et al. (2005). "A fine-scale map of recombination rates and hotspots across the human genome." Science **310**(5746): 321-324.

Nei, M. and T. Gojobori (1986). "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." Mol Biol Evol **3**(5): 418-426.

Norton, H. L., R. A. Kittles, et al. (2007). "Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians." Mol Biol Evol **24**(3): 710-722.

Parra, E. J. (2007). "Human pigmentation variation: evolution, genetic basis, and implications for public health." Am J Phys Anthropol **Suppl 45**: 85-105.

Petes, T. D. (2001). "Meiotic recombination hot spots and cold spots." Nat Rev Genet **2**(5): 360-369.

Pickrell, J. K., G. Coop, et al. (2009). "Signals of recent positive selection in a worldwide sample of human populations." Genome Res **19**(5): 826-837.

Poulter, M., E. Hollox, et al. (2003). "The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans." Ann Hum Genet **67**(Pt 4): 298-311.

Pritchard, J. K., J. K. Pickrell, et al. (2010). "The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation." Curr Biol **20**(4): R208-215.

Relethford, J. H. (2002). "Apportionment of global human genetic diversity based on craniometrics and skin color." Am J Phys Anthropol **118**(4): 393-398.

Ruwende, C., S. C. Khoo, et al. (1995). "Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria." Nature **376**(6537): 246-249.

Sabeti, P. C., D. E. Reich, et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832-837.

Sabeti, P. C., S. F. Schaffner, et al. (2006). "Positive natural selection in the human lineage." Science **312**(5780): 1614-1620.

Sabeti, P. C., P. Varilly, et al. (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **449**(7164): 913-918.

Simonson, T. S., D. A. McClain, et al. (2012). "Genetic determinants of Tibetan high-altitude adaptation." Hum Genet **131**(4): 527-533.

Simonson, T. S., Y. Yang, et al. (2010). "Genetic evidence for high-altitude adaptation in Tibet." Science **329**(5987): 72-75.

Simoons, F. J. (1970). "Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. II. A culture historical hypothesis." Am J Dig Dis **15**(8): 695-710.

Smith, J. M. and J. Haigh (2007). "The hitch-hiking effect of a favourable gene." Genet Res **89**(5-6): 391-403.

Sturm, R. A. (2009). "Molecular genetics of human pigmentation diversity." Hum Mol Genet **18**(R1): R9-17.

Zhang, C., D. K. Bailey, et al. (2006). "A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations." Bioinformatics **22**(17): 2122-2128.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-595.

Tishkoff, S. A., F. A. Reed, et al. (2007). "Convergent adaptation of human lactase persistence in Africa and Europe." Nat Genet **39**(1): 31-40.

Tishkoff, S. A. and B. C. Verrelli (2003). "Patterns of human genetic diversity: implications for human evolutionary history and disease." Annu Rev Genomics Hum Genet **4**: 293-340.

Wang, B., Y. B. Zhang, et al. (2011). "On the origin of Tibetans and their genetic basis in adapting high-altitude environments." PLoS One **6**(2): e17002.

Wang, N., J. M. Akey, et al. (2002). "Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation." Am J Hum Genet **71**(5): 1227-1234.

Voight, B. F., S. Kudaravalli, et al. (2006). "A map of recent positive selection in the human genome." PLoS Biol **4**(3): e72.

Wu, D. D., W. Jin, et al. (2010). "Evidence for positive selection on the Osteogenin (BMP3) gene in human populations." PLoS One **5**(6): e10959.

Yu, F., P. C. Sabeti, et al. (2005). "Positive selection of a pre-expansion CAG repeat of the human SCA2 gene." PLoS Genet **1**(3): e41.

B) Book

Jobling, M. A., M. Hurles, et al. (2004). Human evolutionary genetics : origins, peoples & disease. New York, Garland Science.

# USED WEB ADDRESSES

http://bioinfo.ut.ee/HAD/

http:// ensembl.org/

http://.espcr.org/micemut/

http://hapmap.ncbi.nlm.nih.gov/

http://hgdp.uchicago.edu/

http://pngu.mgh.harvard.edu/purcell/plink/

# ACKNOWLEDGEMENTS

I would like to express gratitude to my supervisors Georgi Hudjašov and Ene Metspalu for all the support, help and experience they have shared with me.

# SUPPLEMENTARY DATA

Supplementary Table 1 including list of samples used in current study could be downloaded from the following link:

http://evolutsioon.ebc.ee/lida/STable_1.xlsx

Supplementary Table 2. Frequency of derived alleles in 200-kb genomic window containing *SLC24A5* gene in studied human population sample. Derived alleles with 100% or 0% frequency are shown in bold. Herein 0% derived allele frequency points to the complete fixation of ancestral allele. Derived alleles were reconstructed using Human Allele Database (HAD), http://bioinfo.ut.ee/HAD/ (Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu).

| rs identifier | Derived allele | Africa | America | Asia | Europe | India | Oceania |
|---|---|---|---|---|---|---|---|
| rs2433354 | C | 0,40 | 0,61 | 0,53 | **1,00** | 0,74 | 0,20 |
| rs2433356 | G | 0,56 | 0,64 | 0,94 | **1,00** | 0,94 | 0,96 |
| rs2675347 | A | 0,40 | 0,61 | 0,53 | 0,99 | 0,74 | 0,20 |
| rs2675348 | A | 0,39 | 0,61 | 0,53 | **1,00** | 0,74 | 0,20 |
| rs8040016 | T | 0,02 | 0,36 | 0,06 | **0,00** | 0,01 | 0,04 |
| rs3736482 | C | 0,72 | 0,64 | 0,94 | **1,00** | 0,94 | 0,96 |
| rs9652449 | G | 0,87 | 0,64 | 0,94 | **1,00** | 0,99 | 0,96 |
| rs8037482 | G | 0,38 | 0,03 | 0,42 | **0,00** | 0,21 | 0,70 |
| rs1878186 | T | 0,36 | 0,97 | 0,58 | **1,00** | 0,75 | 0,27 |
| rs9920281 | A | 0,34 | 0,62 | 0,52 | **1,00** | 0,74 | 0,21 |
| rs12907018 | G | 0,38 | 0,41 | 0,23 | 0,91 | 0,67 | 0,02 |
| rs12912107 | A | 0,30 | 0,48 | 0,27 | 0,81 | 0,65 | 0,16 |
| rs3784614 | T | 0,15 | 0,05 | 0,41 | 0,01 | 0,10 | 0,11 |
| rs11633336 | G | 0,33 | 0,48 | 0,33 | 0,82 | 0,67 | 0,18 |
| rs2279366 | A | 0,50 | 0,45 | 0,50 | 0,17 | 0,36 | 0,80 |
| rs8032420 | C | 0,27 | 0,55 | 0,56 | 0,82 | 0,70 | 0,21 |
| rs12593807 | C | 0,42 | 0,45 | 0,44 | 0,14 | 0,28 | 0,73 |
| rs1878187 | A | 0,69 | **1,00** | 0,94 | 0,96 | 0,91 | 0,93 |
| rs8040834 | C | 0,15 | 0,14 | 0,33 | **0,00** | 0,13 | 0,18 |
| rs8039702 | T | 0,38 | 0,45 | 0,46 | 0,13 | 0,29 | 0,73 |
| rs6493317 | T | 0,38 | 0,55 | 0,49 | 0,82 | 0,62 | 0,20 |
| rs6493318 | C | 0,62 | 0,45 | 0,51 | 0,18 | 0,38 | 0,80 |
| rs7179027 | A | 0,52 | 0,55 | 0,54 | 0,82 | 0,68 | 0,21 |
| rs17350938 | G | 0,08 | 0,32 | 0,13 | 0,13 | 0,17 | 0,57 |
| rs8025278 | G | 0,56 | 0,26 | 0,37 | 0,04 | 0,19 | 0,14 |
| rs964611 | T | 0,09 | 0,32 | 0,13 | 0,14 | 0,19 | 0,57 |
| rs7168752 | C | 0,09 | 0,32 | 0,13 | 0,14 | 0,19 | 0,57 |
| rs2413891 | G | 0,46 | 0,74 | 0,60 | 0,97 | 0,77 | 0,86 |