University of Cambridge
Division of Biological Anthropology

Genetic insights about adaptation to high altitude in the Caucasus region

Student number: 0402D
Word count: 9730

Abstract

My study is aimed to investigate genetic adaptation in the Caucasus in populations of moderately high altitude. So far extreme high altitude adaptation has been mainly studied in Andean and Tibetan populations, where several selected genes related to hypoxia have been identified. First insights into genetic high-altitude adaption for Caucasus people were provided in a study focusing on 15 high-altitude adaptation candidate genes. While this previous candidate gene-approach was restricted to already known high altitude genes, my study on genome-wide data has the potential to reveal new selected genes including those in response to high altitude. Published genome-wide Single nucleotide polymorphism (SNP) data was obtained from 10 different ethnic groups of the Caucasus. Although none of the ten populations were sampled from higher than 2000 m altitude the search for high altitude genes was motivated by the long term population histories in moderately high altitude regions. The data set was screened for recent signs of selection using the haplotype-homozygosity based test iHS. Regions with significantly high empirical iHS values were further screened for enriched pathways or individual genes that are related to hypoxia. In Balkars, the gene *CFTR* stood out. In Abkhazians, genes involved in the lipid metabolism PPAR pathway and *Caveolin 1* and *2* showed signs of selection. The PPAR pathway is connected to the HIF-pathway and was previously detected among Tibetans' high altitude adaptation signals. *Caveolin 1* is associated to nitric oxide synthesis regulating vasodilation. Besides high altitude associated loci, the iHS results were screened for the presence of longevity associated genes. Two genes *RPA3* and *SMG6* related to DNA replication and telomere activity could be identified in Abkhazians. However, it is uncertain if Abkhazians have indeed a higher life expectancy. In sum, the study revealed several genes in two Caucasus populations with signs of selection that are likely to have evolved in response to hypoxia. Furthermore, two genes with signs of selection were detected among longevity related genes.

Table of content

1 Introduction

1.1 High altitude adaptation

1.1.1 How people adapted to high altitude

Hypoxia as a strong selective pressure
High-altitude environments serve as an ideal natural laboratory for studying adaptation. With rising height the oxygen content in the air decreases. For instance, at 2000 m elevation, every breath of air contains less than 80 % of the oxygen molecules compared to the same breath at sea level (Fig. 1) While other environmental stresses on the world such as temperature, UV radiation, and diet could be buffered by cultural changes, hypoxic environment is an unavoidable and continuous pressure for humans living in high altitude regions. To find out about the stress that hypoxia puts on humans it is best to start with how lowlanders react in a hypoxic environment. In the first place heart and breath rate are increased to compensate for the low oxygen content in the air. In a later stage, acclimatization takes place. This is expressed by an increase of the Hemoglobin (Hb) concentration improving the oxygen carrying capacity of the blood. However, chronic exposure to hypoxia can lead to chronic mountain sicknesses characterized by an excessive red blood cell production, thereby causing an increase in blood viscosity and an uneven blood flow through the lungs (Monge and Leon-Velarde 1991; Cheviron and Brumfield 2012). Additionally, the body response to hypoxia in the lung is vasoconstriction of pulmonary vessels. In this way hypoxic tissue tries to compensate the lack of oxygen with a better blood supply. However, the resulting pulmonary hypertension can lead to severe complications in the lung. A very serious life-threatening form is the high altitude pulmonary edema (HAPE) (Murray, Insel et al. 2006). Chronic hypoxia furthermore causes fetal intrauterine growth restriction in pregnant women (Moore 2003). All these reactions demonstrate that hypoxia has a very strong impact on the organism.
The physiological responses to hypoxia expressed by lowlanders might help the organism in the first place. But as seen in the example of HAPE, there are cases in which reactions of the body towards hypoxia are disadvantageous and rather counter productive. Additionally, all these responses are energy consuming and for a chronic exposure to hypoxia not optimal. Indeed, highland natives physiology is completely different in high altitude compared to lowlanders in the same altitude. The adaptation to high altitude of native highlanders will be considered in the following section.

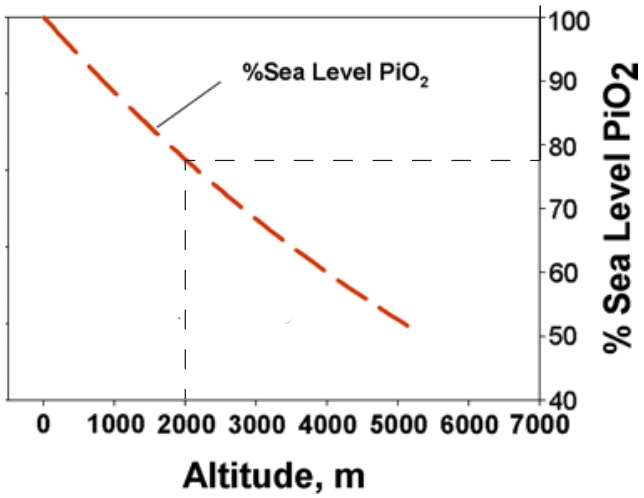

Figure 1: **Oxygen levels decrease with increasing altitude** The percentage of partial pressure of inspired oxygen (PiO_2) at sea level is plotted against the altitude. At 2000 m, PiO_2 has fallen to less than 80 %. Plot adopted from (Beall 2007) and modified with Adobe Illustrator.

High altitude adaptation in Tibetans, Ethiopians and Andeans

As discussed, hypoxia is an important stress factor for an organism. Distinctive features special for native highlanders suggest a genetic adaptation to high altitude. A myriad of studies have been carried out proving this assumption and revealing genes that were selected in response to high altitude. The three populations in which high altitude adaptations were mainly studied are Tibetans, Ethiopians and Andeans.

Tibetans adaptation, for instance, is reflected in several phenotypic traits. They don't develop chronic mountain sickness (Monge and Leon-Velarde 1991). Furthermore, pulmonary vascular structure, lack of hypoxic pulmonary vasoconstriction, and a lower incidence of reduced birth weight are all traits that are likely to have evolved because of high altitude stress (Beall 2007). The most striking fact is that they maintain sea-level Hb concentration in a height of up to 4000 meters (Cheviron and Brumfield 2012).

A key role in genetic high altitude adaptation is the HIF (hypoxia inducible factor) pathway. HIF is a transcription factor which is activated in hypoxic conditions and triggers the transcription of at least 70 genes related to hypoxia response (Beall 2007). One of them is for instance *EPO*, which in turn induces red blood cell production (Simonson, Yang et al. 2010). Genes involved in the HIF pathway have found to be the target of selection in all studied high altitude populations Tibetans, Andeans and Ethiopians (Bigham, Mao et al. 2009); Scheinfeldt, Soi et al. 2012). In the case of Tibetans, it is believed that alterations in genes involved in the HIF pathway (for instance *ELGN1* and *EPAS1)* are the cause for the low HB concentration as the hypoxia induced erythropoietic response is blunted (Cheviron and Brumfield 2012). There are a wide variety of other detected genes in the three populations, which would be beyond the scope of this introduction. However, it is noteworthy that each of these populations took different pathways to adapt to high altitude. For instance, the Hb concentration in Andeans is increased, as opposed to the decrease of Hb concentration in Tibetans. The three populations have evidently adapted separately to high altitude adaptation, originating from different starting populations. There are therefore different genes selected in the three populations. However, as it is the same selection pressure, common pattern can be observed. As in all three populations genes involved in the HIF-pathway were detected.

1.1.2 Caucasus

General information about Caucasus populations

The focus of this study is on the populations of the mountainous region of the Caucasus. Anatomically modern humans appeared in this region at least 42 thousand years ago and there is evidence for continuous human occupation since 10 000 years ago (Adler, Bar-Yosef et al. 2008; Caciagli, Bulayeva et al. 2009). There is further evidence that highland villages might have served as refuges during the last glacial maximum (LGM) (Caciagli, Bulayeva et al. 2009). Many Caucasus ethnic minority groups have a long history of isolation with a high degree of inbreeding, having been endogamous and isolated for centuries (Pagani, Ayub et al. 2012). Furthermore, the Caucasus has undergone through demographic change due to emigration, immigration, internal conflicts and imposed evacuation. For instance, in 1864 most of the Muslim Abkhaz left the Caucasus (Colarusso 1995). Mingrelians, a subethnic group of Georgians were relocated under Stalin from 1937 to 1953. During the Second World War, also Chechens and Balkars had to leave their homelands (Grannes 1991). Other populations, such as the Kuban Nogays (K Nogays) emigrated to the Caucasus from the Pontocaspian steppes in late 18th-early 19th century (Kolga, Tonurist et al. 2001).

Figure 2: **Map of the Caucasus region** The physical map shows the Caucasus mountains, with the East European plain in the North and the Middle East in the South. The highest points are the summits of Elbrus (5644 m) and of Kazbek (5033 m). Map adapted from www.mapsof.net

High altitude adaptation in Caucasus populations

The Caucasus Mountains reach 5644 and 5033 m at their highest points (Fig. 2). Like the Andeans or the Himalaya, this region can be used to study how people adapted to high altitude. Only a few studies exist that investigate high altitude adaptation in Caucasus populations. This is due to the fact that the requirements for studying high altitude adaptation (long residence and high elevation) are less clear than in the other studied populations. While Tibetans live at elevation exceeding 4000 m, the highest residences in the Caucausus are approximately 2300 meters high (Yi, Liang et al. 2010). Additionally, as discussed in the previous section, the origins and migrations of some

ethnic groups are complex. However, there is some genetic evidence of high altitude adaptation in the Caucasus populations.
In 1987, a variant Hb alpha subunit sequence was discovered in Dagestan individuals (Lacombe, Arous et al. 1987). A recent study with a candidate gene approach studied Dagestan highlanders living in 2000 m of altitude (Pagani, Ayub et al. 2012). Two genes *HIF1A* and *ELGN1* showed signatures of positive selection. *ELGN1* regulates HIF and is associated with a low Hb concentration in Tibetans. *HIF1A* is related with an improvement in oxygen metabolism. To conclude, there is some genetic evidence for high altitude adaptation. However, considering the limited scale of genetic variation assessed by these former studies, there are potentially more genes that have been under selection in response to high altitude in the Caucasus.

1.2 Longevity

The Caucasus data set used in this study evidently has potential to reveal other adaptation except from that to high altitude. Cases of extreme longevity in some Caucasus populations have been reported in the literature, although often these are not confirmed by concrete evidence. Hence, as a side project in this study, the results of selection scanning on the genome-wide data set was also subjected to a screen for genes related to longevity.

1.2.1 Longevity in Abkhazians

In the early eighties, many reports for extreme longevity in certain isolated regions were published. Amongst these regions were some villages in the Caucasus, especially in Abkhazia, where an unusual high number of centenarians lived (Kyucharyants 1974) (Fig. 3). In a 5-year multidisciplinary study of longevity in Georgians it was found that the index of longevity (N90 + /N60 +) for some villages in Abkhazia is much higher than the average level for the rural population of the whole region (Lelashvili and Dalakishvili 1984).
However, a revaluation of the data published in 1984 questioned the longevity reports published until this point (Palmore 1984). The revaluation, based on a survey of 4 villages in Abkhazia, showed that only 0.3 % of the Abkhazians were over 90 years old, a number not higher than in the USA. Claims of extreme ages were presumably exaggerated for political, social and cultural reasons (Kyucharyants 1974). In 1917, after the Russian revolution, a new system of registration was introduced in rural Abkhazia. Here it is very likely that incorrect methods in age registration lead to awarding additional years to some persons to enhance their social status (Bennett and Garson 1986). On balance, the overall opinion in literature today is that there is no special longevity in the Caucasus, or at least there is not enough evidence to justify further investigation about longevity in these populations. Given the existing genome-wide data, however, it is possible without any further investments to test whether genes related with longevity stand out in the Caucasus populations.

Figure 3: **A 138-year old Abkhazian woman with her 85-year old son in 1974** In a revaluation 10 years later, it became clear that there was no documented evidence and the extreme ages in Abkhazians villages had presumably been exaggerated. Picture source: (Kyucharyants 1974)

1.2.2 Longevity and genetics

Even if the Abkhazians (against the conventional opinion) have a higher life expectancy, one cannot automatically assume a genetic cause for this phenotypic trait. One has first to find out about the heritability of longevity. It could be that the environment, the way of life, the diet exclusively decides about how long a human being lives. Longevity is especially difficult to study because there are a lot of different factors that cause a premature death. However there is evidence that, at least partially, suggest a genetic influence for duration of the life span.

First of all, there is a lot of evidence in model organisms. For instance, a single point mutation in mice delays the rate of ageing by about 30 % and a single mutation in the nematode *C. elegans* doubles life span (Migliaccio, Giorgio et al. 1999) and (Kenyon, Chang et al. 1993). Studies investigating humans are more difficult to carry out. However, a number of twin studies gave some evidence for heritability. It was shown that approximately 25 % in the variation of human lifespan is explained by genetic factors (Schoenmaker, de Craen et al. 2005).

There are also several premature ageing syndromes all caused by genetic mutations that suggest a genetic basis for ageing. Premature ageing syndromes are characterized by typical ageing symptoms that occur in a much earlier age. The most famous example for such a disorder is the Werner syndrome (WS). WS patients have a median life expectancy of 47-54 years. In the second decade of life, they develop signs and pathologies which resemble many aspects of normal human ageing (Coppedè 2012). Amongst the symptoms are greying of hair, thickening of the skin, diabetes mellitus, cancer and atherosclerosis. The genetic origins of WS are mutations in the Werner syndrome protein (WRNp), which causes telomere attrition and genomic instability. While disorders in ageing are well studied, only few association studies exist investigating centenarians. Some evidence suggests that there are special alterations found in Human leukocytes antigen HLA-DR, Apolipoprotein E, Angiotensin-converting enzyme ACE (Christensen and Vaupel 1996). However, these findings have to be reproduced and are not yet reliable.

In summary, there is a lot of evidence that the longevity trait is partially heritable and genetic disorders without any environmental influence can cause premature ageing. On the other hand, environmental and public-health factors are without a doubt a similar or an even stronger determinant of longevity than genetic factors are. The best evidence for this is the large increase of life expectancy in the last hundred years in the industrialized world, which would prove difficult to explain exclusively with genetic changes (Christensen and Vaupel 1996). So, when discussing possible genetic causes for the longevity in Abkhazians, potential environmental factors that could also contribute to longevity also have to be taken into consideration.

1.2.3 Telomeres

Understanding the mechanism of telomeres and telomere attrition is central in understanding the physiology of ageing and longevity. Telomeres are non-coding, repetitive sequences at the end of DNA. Together with proteins they protect the DNA from deterioration and from fusion with other chromosomes (Sampedro Camarena, Cano Serral et al. 2007). During every cell division, they shorten if not restored by telomerase activity. If it comes to a critical length, the cell stops dividing (replicative senescence). Telomeres therefore have two important functions. On the one hand, they ensure that coding sequences don't get lost, on the other hand they limit cell replicative potential (Heidinger, Blount et al. 2012).
There are numerous studies investigating the correlation between life span and telomere length. For instance, in a study measuring telomere length in zebra finches, it was found that individuals living longest had relatively long telomeres at all life stages (Heidinger, Blount et al. 2012). Even though many studies have been carried out, the actual mechanism of telomeres behind the ageing process of an organism remains unclear and controversial.

1.3 Background of test statistics

1.3.1 Phasing

Tests of haplotype homozygosity, including iHS, require phased data as their input. Most genotyping methods do not determine directly whether two particular alleles in adjacent SNPs have been derived from the same or different parental chromosomes unless they are phased (Browning and Browning 2011). In unphased data, every diploid binary SNP locus can carry two alleles. For instance, two loci are expressed as follows: A/C and G/T. The two possible haplotypes that, reflect the combined inheritance of the genetic loci from the same parental chromosome, are consistent with the genotype data are: AG and CT or AT and CG. The process of phasing is needed to infer from the unphased genotype data the phased haplotypes. The resulting inference of phased data, considering information from other genetic loci in the vicinity of the two, and frequency data from the relevant population, would for instance be: AG and CT. The most powerful method of phase inference relies on mother-father-child trios where the phase of one of the parents can be resolved from the information of the child and the other parent but such trio-data are not commonly available for populations of interest. Direct determination of phase of molecular markers through cloning or cell sorting technologies can be either extremely time consuming or expensive or both.
Computational methods that provide approximate phasing are different for related and unrelated individuals. For unrelated individuals, they can model haplotype frequencies in a set of individuals to estimate the most probable haplotype configuration. The software BEAGLE is used in this study for phasing genome-wide genotype data. It is

based on the hidden Markov Model and its accuracy increases with sample size (for more details, see (Browning and Browning 2007)). One has to be aware of the fact that computational phasing is an approximation and doesn't guarantee 100 % accuracy.

1.3.2 iHS (integrated haplotype score)

long-range haplotype as sign of positive selection

iHS is a haplotype based test designed to identify genes that were recently target of selection. The test compares the length of haplotype homozygosity for all SNPs in the genome-wide data for their ancestral and derived allele. When a new derived allele is generated by mutation, it will, at first, have a very low frequency in the population (essentially 1/2N, where N is the populations size) and be linked with any other marker on the chromosome of its origin. Over time the linkage will be continuously lost with other loci on the chromosome due to recombination, yet the linkage survives for many generations with SNPs that are near to it (Fry, Trafford et al. 2006). This is called a linkage disequilibrium (LD), defined as a non-random association of alleles at two or more loci (Jobling, Hurles et al. 2004). Normally, at the birth of any new allele the surrounding LD is high but its frequency low (Fig. 4). If the new allele, however, is advantageous, its frequency can increase very quickly. In this short time period, recombination doesn't have time to break down the haplotype. As a result, the selected allele sits on a long haplotype with low diversity (called a long-range LD) (Fig. 4). Selected alleles are therefore characterized by relatively high frequency paired with a long-range LD. Conversely, neutral alleles have a short-range LD but might have the same frequency due to genetic drift.

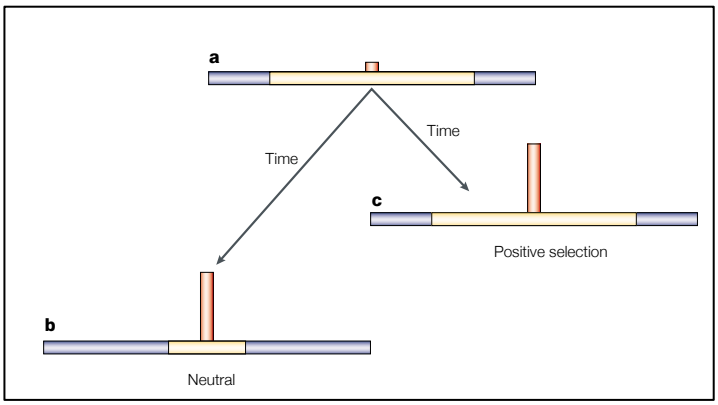

Figure 4: Scheme of how LD changes under neutral and positive selection a) a new allele occurs and starts with a low frequency (indicated by the height of the orange bar b) Under neutral selection, the range of LD is reduced by recombination events; frequency of the allele increases due to genetic drift c) In case of positive selection, frequency of the allele increases very quickly. The increase occurs faster than recombination events can reduce the range of LD.

Test statistics

EHH (extended haplotype homozygosity) is the primary statistic by which iHS is estimated. EHH measures the decay of identity as a function of the distance, starting from the "core" allele (Pickrell, Coop et al. 2009). At the core allele, if the haplotype homozygosity is set to 1 then with increasing distance, the homozygosity decays, until EHH is 0. The EHH decay for a recently selected derived allele is expected to be less strong than for a neutral allele because of its long-range LD (Fig. 5). One has to take into account that for each region of the genome the recombination rate differs. Therefore, to capture those alleles that have indeed a more

extended homozygosity, every EHH curve is compared to the corresponding curve of the ancestral allele. The first step is to calculate the IHH (integrated EHH - area under the curve in both directions from the core allele for ancestral (IHH_A) and derived allele (IHH_D) respectively). The unstandardized iHS is then defined as follows:

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

It has further to be taken into account that iHH for SNPs with higher frequency are expected to have a lower value. The final iHS statistic is standardized so that iHS values from alleles with different allele frequency within a genome are comparable. When iHS is approximately 0, derived and ancestral allele have the same EHH decay and one can assume that the allele was under neutral selection. Large negative iHS values for a derived allele indicate an extended haplotype compared to the haplotype of the ancestral allele. Therefore, large negative iHS values are a sign of potential positive selection. These sweeps can produce alleles with large positive iHS in their surroundings. Therefore, negative and positive iHS values are considered in the test statistics to increase the possibility of sweep detection.
It is important to note, that selected alleles that reached high frequency or fixation have lost their long-range LD. iHS statistics therefore can only detect intermediate frequency alleles which have undergone recent strong selection and where the extended haplotype hasn't been broken down yet (Voight, Kudaravalli et al. 2006). Older selective events with alleles that reached high frequency or fixation therefore cannot be detected. iHS can have a power up to 40% to detect signals of positive selection in cases where the selected allele has a frequency 40-60%, the case of soft sweep (Pickrell et al. 2009).

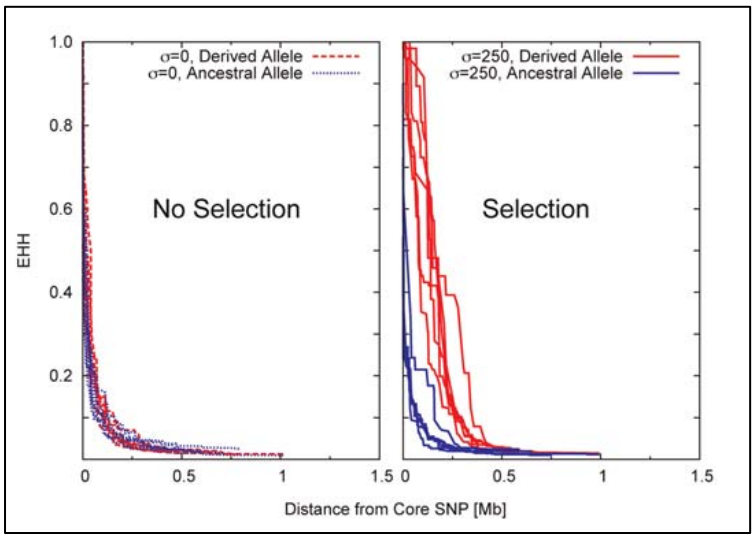

Figure 5: **EEH decay of ancestral and derived allele on simulated data for an allele at frequency 0.5**. The distance from the core allele is plotted on the horizontal axis, the EHH value on the vertical axis. In case of no selection, derived and ancestral allele EHH decay are similar (left panel). If the derived allele shows recent signs of selection, its EHH curve decreases less quick than the one from the ancestral allele (right panel). Signs of positive selections can be detected by comparing the area under derived and ancestral alleles curves. Source of the plot: (Voight, Kudaravalli et al. 2006)

1.4 Aim of this study

There is genetic evidence for high altitude adaptation in Ethiopians, Tibetans and Andeans. Recently, it could also be shown that there is genetic high altitude adaptation in a population of the Caucasus. The aim of this study was to investigate high altitude adaptation in Caucasus population more broadly and in more detail. iHS statistics were carried out on a haplotype data set from major ethnic groups of the Caucasus region. The results were screened for signs of hypoxia response. Furthermore, Caucasus populations are also known, although with some level of controversy, for their longevity. Therefore, iHS results were also screened for signs of longevity.

2 Methods

2.1 DNA samples

The genome-wide genotype data for the Caucasus populations were obtained from a published study that analyzed the relations between the individual Caucasus populations but without carrying out any selection scans on the data (Yunusbayev, Metspalu et al. 2011). DNA samples had been collected from major ethnic groups of the Caucasus region (Fig. 6). Physiological parameters of the donors were not known. DNA samples had been genotyped with the Illumina 610 K SNP array. The data for the support groups used in this study (Han Chinese, French, Lebanese, Syrians, Turkmen) are from another published source, the Human Genome Diversity Project (Li, Absher et al. 2008).

Altitude information for every sampling point was inferred from latitude and longitude data via gpsvisualizer (http://www.gpsvisualizer.com/elevation) (Tab. 1 and Tab. 2). It is important to note that locations of the sampling points give only residence information for the past few generations or even less.

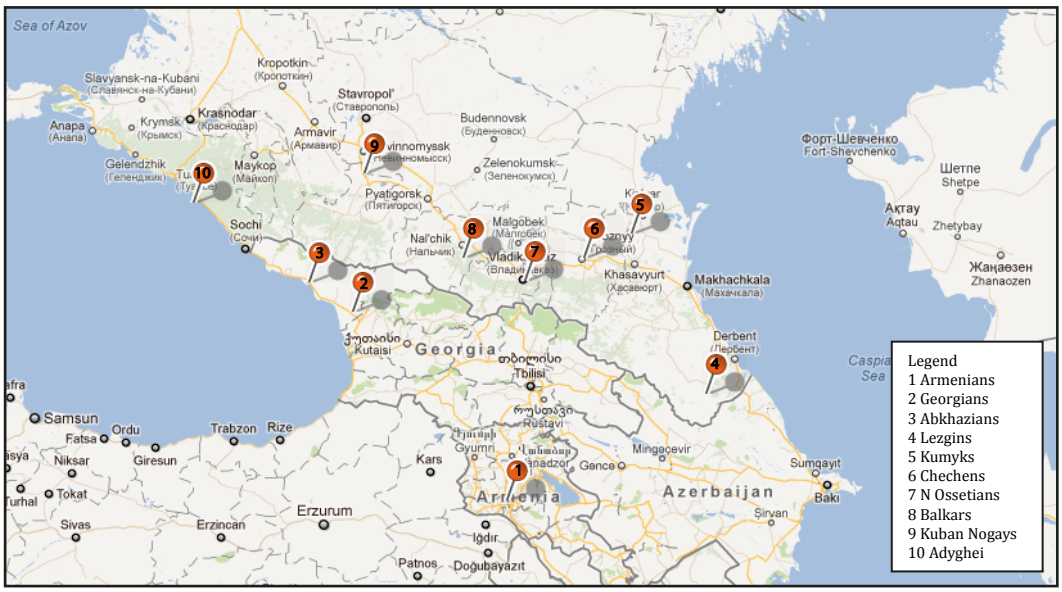

Figure 6: **Geographical map with the populations included in this study** Pins indicate the DNA sampling points from the different ethnic groups of the Caucasus that were analysed in this study (the map was created using www.sketchmap.co.uk)

Table 1 List of populations used in this study

population	sample size	elevation
Armenians	35	1232
Georgians	20	56
Abkhazians	20	15
Lezgins	18	1341
Kumyks	14	5
Chechens	20	130
North Ossetians	15	686
Balkars	20	715
Kuban Nogays	16	426
Adyghei	17	5

Table 2 Summary of control population outside the Caucasus

population	sample size	elevation
French	34	168
Syrians	16	525
Lebanese	8	1070

2.2 PCA and ADMIXTURE

PCA (principal component analysis) and ADMIXTURE were carried out in order to get population structure and to know which populations with sample size < 20 were to match for iHS calculations. Principal components (PC) on a) all Caucasus populations, control populations and major ethnics groups and b) Caucasus populations and control populations were calculated using Eigenstrat (Price, Patterson et al. 2006). Results of the calculations were presented using Excel Office 2011.

2.3 iHS calculation

iHS calculation can only be applied on phased data. Phasing was carried out using BEAGLE software package (Browning and Browning 2007). All non-African populations were phased together, using CEU as a reference (CEU: samples collected in Utah, with North and South European ancestry). PLINK software was used to filter the combined dataset to include SNPs only on the 22 autosomal chromosomes and for which ancestral SNP was known (Purcell, Neale et al. 2007). Ancestral alleles were inferred from Ensembl database. In the end, 528,741 SNPs were used.

iHS was calculated as in Voight et al (2006). The genome was split into non-overlapping segments of 200 kb windows. For each window, the iHS value for every SNP with minor allele frequency of 5 % was calculated. In order to be able to rank the windows, following approach was made: the windows were categorized according to how many bins they have. The value of 0 bin corresponds to 0-20 SNPs, 1 corresponds to 20-40 and so on. Within each category, the windows were ranked according to their cut value. The cut value is the percentage of the SNPs that have an iHS value above 2 in the given window. The p-value then is calculated in accordance with the rank. The windows that are at top of the rank have for instance a p-value of 0, as they are the most significant ones. Information and corresponding explanations are summarized in Table 3.
The windows with a bin value of 0 were removed. All the other windows were ranked with increasing p-values. Within the same p-Value the windows were ranked with decreasing cut values. For each Caucasus population, the windows that are not present in the control groups (French and Lebanese_Syrians) were filtered. The first 1 % of the Caucasian populations respectively was filtered with the first 5 % of the supports group (Fig. 7). Genes for the resulting windows were distinguished using BioMart. These gene lists were then analysed for enrichment.

Table 3: Summary of information provided for each window.

Chromosome and position	location of the window
Cut	percentage of the SNPs that have an iHS value above 2 in the given window $\frac{\#SNPs\,|\,(iHS > |2|)}{\#SNPs}$
Mean	the average iHS value of all the SNPs in the given window
Bin	number of bins, that are present in the given window; bin=0 ⇒SNP (0-20)
p-Value	for every window with a certain bin value, an order is made ranked depending on the cut. The top of this rank gets a p-Value of 0 (most significant hit)

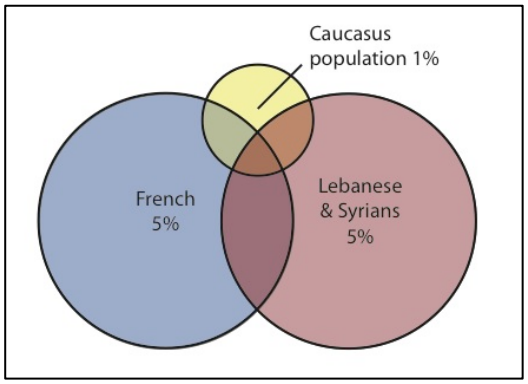

Figure 7: **Schema of iHS calculation analysis** The yellow area which is neither present in French nor in Lebanese_Syrians is used for further gene enrichment analysis (circles not to scale).

2.4 DAVID gene enrichment

In order to assign biological meaning to the calculated gene lists, DAVID gene enrichment software (DAVID v6.7) was used (Huang, Sherman et al. 2008; Huang, Sherman et al. 2009). It was used to search for genes with possible correlation to high altitude adaptation in both functional annotation clustering and term clustering. Modified Fisher's exact test was used to determine the significance of gene-term enrichment. The general cut-off point is 0.01. However, also term enrichments with higher p-value were considered.

2.5 Candidate gene approach

Gene enrichment can only detect several genes assigned to the same biological function. To detect individual altered genes, a candidate gene approach was developed. Different lists of genes related to high altitude were used to screen for individual significant genes. One list was taken from the study investigating high altitude adaptation in Dagestani (Pagani, Ayub et al. 2012). Two lists were obtained from two studies analysing the high altitude adaptation of Ethiopians and Tibetans (Simonson, Yang et al. 2010; Scheinfeldt, Soi et al. 2012). Furthermore, a gene list was created with all the genes that were annotated by the GO (gene ontology) terms "response to hypoxia", "oxygen transport" and "oxygen homeostasis" (listed in supplementary Tab. 2). Apart from the hypoxia list, a gene list for longevity was created. Genes with the GO annotation terms "telomere

maintenance", "senescence" and "cell ageing" were included (listed in supplementary Tab. 3).

2.6 Which genes are common in all the Caucasian populations but not in control groups?

In order to find common significant windows in all Caucasus populations, a less conservative approach was made. In opposite to the first 1 %, that was used for the individual populations, this time the first 5 % were used. All the non-overlap windows of all the Caucasus populations (windows that were not present in the first 5 % of the support populations, but were present in the first 5 % of all the Caucasus populations) were matched in order to get significant windows that are not present in the support groups but are present in all the Caucasus populations.

3 Results

3.1 PCA and Admixture

The PCA plot including Lebanese, Syrians and Turkmen shows a dense clustering for all Caucasus population except for Armenians and K Nogays. As expected control groups are further away form the Caucasus population cluster (Fig. 8). To get a more detailed picture of the Caucasus populations, a second PCA was carried out, excluding the control groups. Here, K Nogays is once again an outlier, while the populations of the South Caucasus (Armenians, Georgians and Abkhazians) cluster together, as expected from their geographical location (Fig. 9). The admixture plot distinguishes clearly between Yoruba, French and Han Chinese. K Nogays shows a very similar distribution to the Turkmen. The populations of the South Caucasus are distinguished by the fact that they don't have a Han Chinese proportion, as opposed to the rest of the Caucasus populations (Fig. 10).

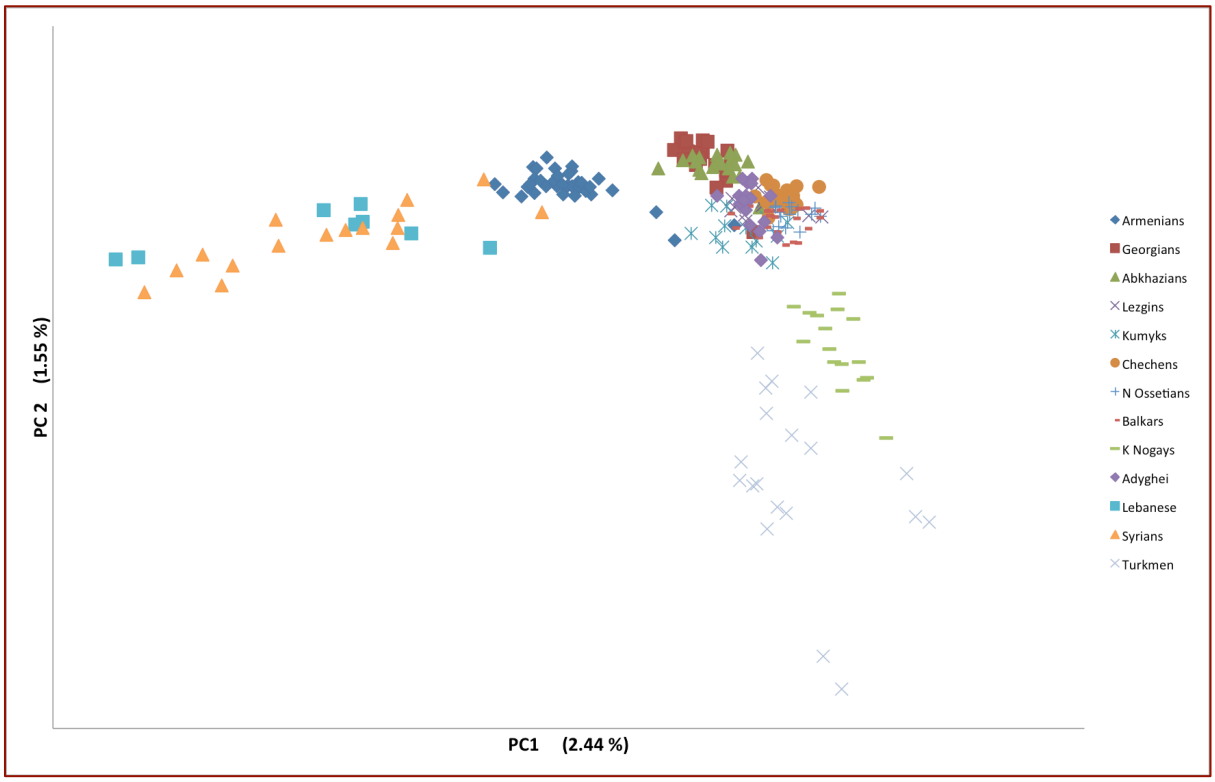

Figure 8: **PCA of the Caucasus populations and the control groups** Plot of the first and second components of PCA of the Caucasus populations, Syrians, Lebanese and Turkmen. Most of the Caucasus populations cluster together as expected from their geographical location. K Nogays are outliers, plotted towards Turkmen. On the other side, Armenians are outliers, approaching Lebanese-Syrians. The percentage in brackets indicates how good the principal component describes the data set. Plot was created using Excel Office 2011.

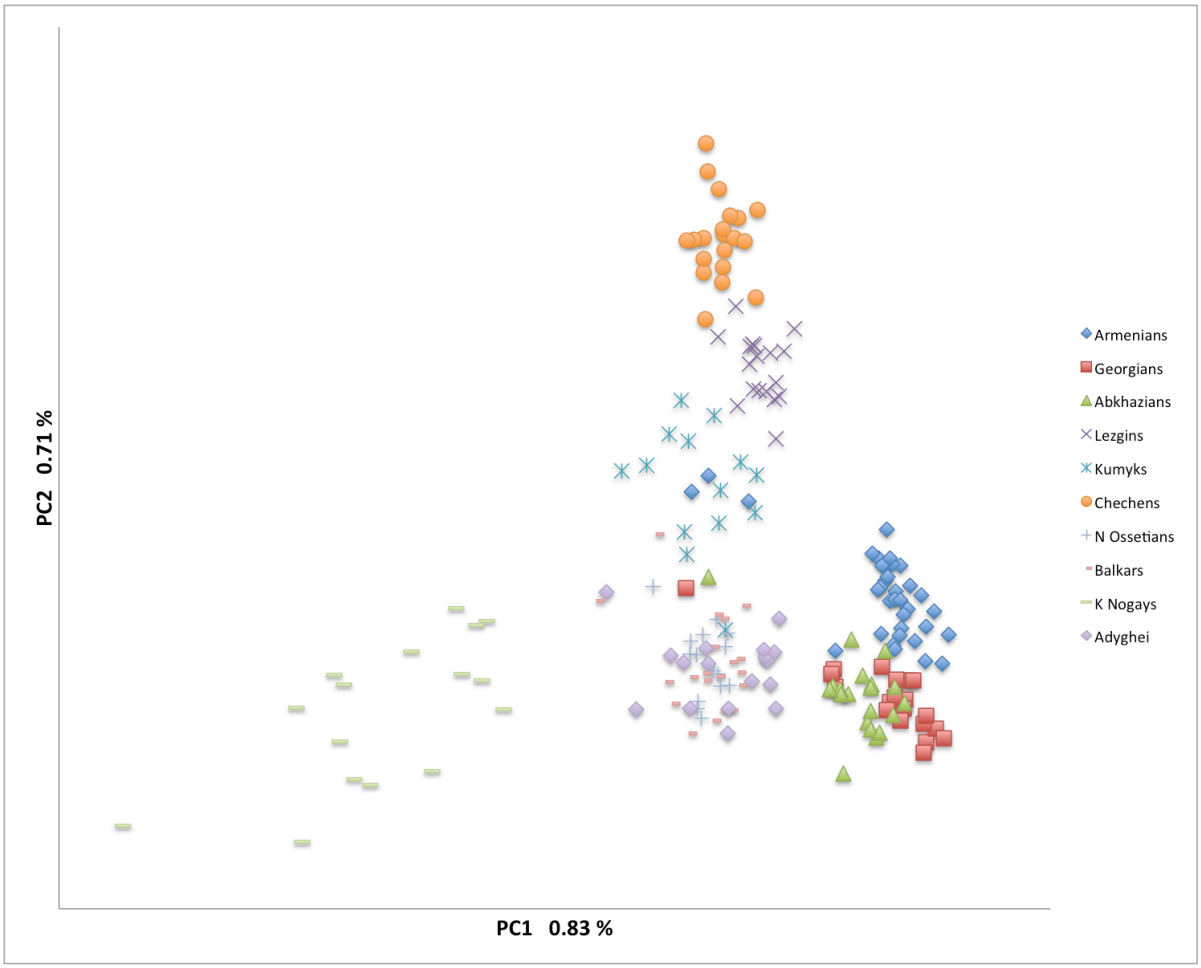

Figure 9: **PCA of Caucasus populations without control populations** Plot of the first and second components of PCA of the Caucasus populations. K Nogays is identified as an outlier population, clustering with no other population. The populations of the South (Armenians, Georgians and Abkhazians) are closely clustered together, while Chechens are further away from most of the Caucasus populations. The individual PC percentage indicates how good the PC describes the data set. Plot was created using Excel Office 2011.

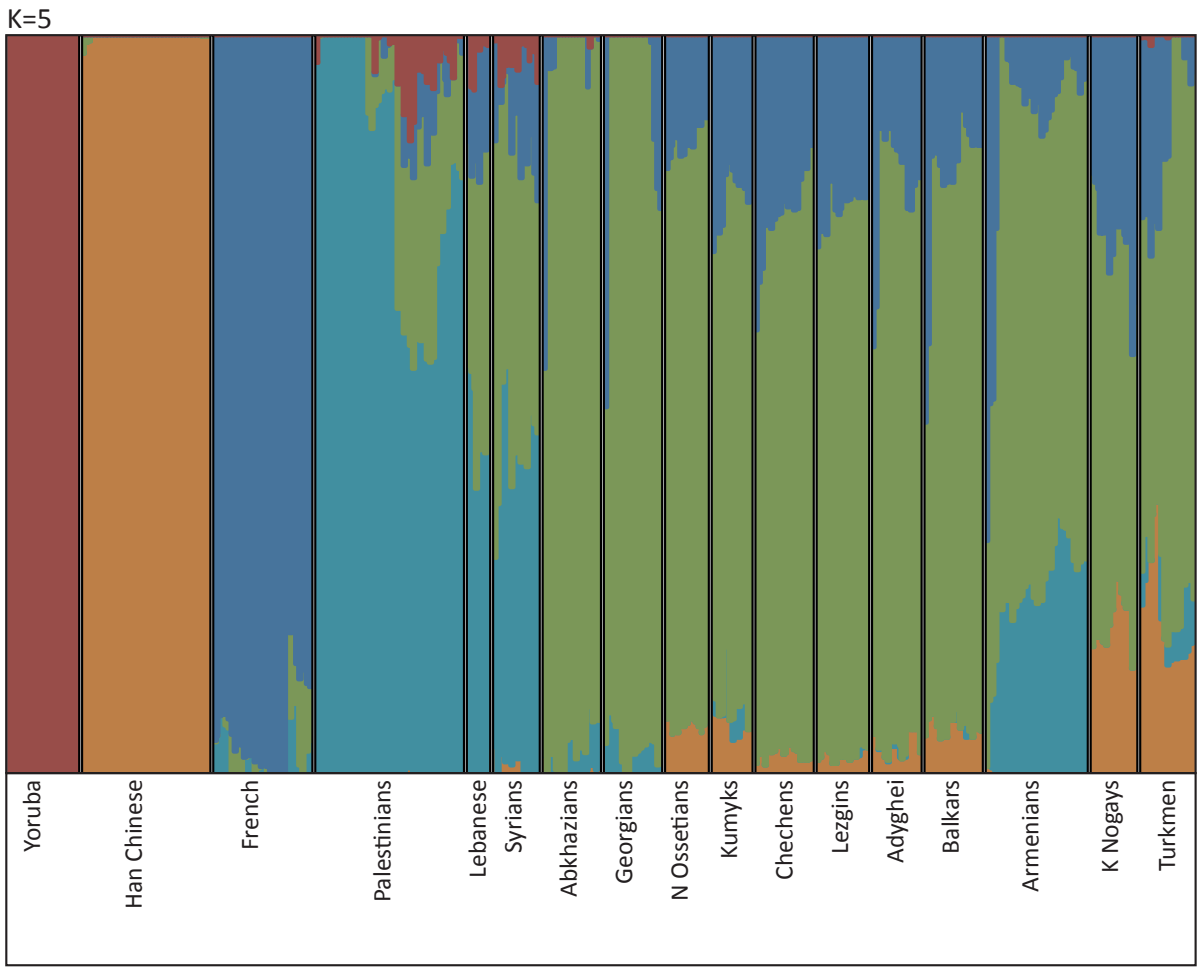

Figure 10: **Admixture of the Caucasus populations, control groups and Han Chinese and Yoruba** Population structure inferred by ADMIXTURE analysis of the genetic data at K=5. Each vertical column represents one individual, divided according to ancestry probabilities proportions. Yoruba, Han Chinese and French can clearly be distinguished from each other. K Nogays have similar ancestry proportions to Turkmen. Armenians proportions are close to those from Syrians and Lebanese. Plot was created using Excel Office 2011.

4 populations were grouped together for iHS calculations

Kumyks, Lezgins, Adyghei, N Ossetians and K Nogays are populations with an insufficient amount of samples for iHS calculations (threshold 20) (Tab. 1). For the control, Syrians has a sample size of 18. PCA and ADMIXTURE results should help to decide which populations to combine as a group. N Ossetians and Adyghei cluster together. Lezgins and N Ossetians form individual clusters. It was decided to give elevation the priority and therefore the two lowlander populations Adyghei and Kumyks were grouped together for iHS calculations. K Nogays has been excluded from further analysis steps as they are outliers. Palestinian samples do not form a homogenous cluster as seen in the Admixture plot (Fig. 10). Indeed, about a fifth of the individuals show the same pattern as Lebanese and Syrians, while the rest of them have their own specific background. Considering the potentially mixed ancestry of the Palestinian sample they were also excluded from further analysis. Lebanese and Syrians have in opposite the same homogenous admixture pattern (Fig. 10). Therefore, it was decided to use Lebanese as an additional control group and group Lebanese and Syrians together. However, it has to be noted that altitude of Lebanese is quite high (1070 m), but Syrians couldn't have used on its own (populations used for iHS calculations are summarized in Table 4).

Table 4 List of populations for iHS calculations

Abkhazians
Armenians
Balkars
Chechens
Georgians
Kumyks_Adyghei
Lezgins_NorthOssetians
French
Lebanese_Syrians

3.2 DAVID gene enrichment

DAVID gene enrichment was employed to find out about optional enriched pathways or annotation clusters in the top 1 % of iHS of every individual Caucasus population excluding genes from the top 5 % of the control groups. Only these populations will be discussed that showed interesting enrichments in terms of high altitude adaptation. Other enrichments and the gene lists used for the enrichment study can be found in the appendix (Supplementary Tab.1 and Supplementary Tab. 4).

3.2.1 Abkhazians

In the first 1 % of Abkhazians, genes were enriched for the PPAR signaling pathway. The three genes enriched are *ACOX1*, *PLPT* and *RXRG*. The p-value for this enrichment is 0.027 (Tab. 5). The p-value here indicates the EASE score, a modified Fisher exact p-value (for more information see (Hosack, Dennis et al. 2003)). The PPAR pathway had already arisen in studies of high altitude adaptation of Tibetans. The possible biological meaning will be discussed in more detail in the next section.

3.2.2 Balkars

In the first 1 % of Balkars, genes were enriched for the term "lung development". The five genes enriched are *ASZ1*, *Cttnb2*, *CFTR*, *MAPK8IP3* and *PROX1*. The p-value for this enrichment is 0.00029 (Tab. 5). The *CFTR* gene was an especially interesting finding because its loss of functionality has such a big impact on the body, leading to mucoviscidose. The aforementioned gene is present in two contiguous windows (window A: 1,170K bp – 1,172K bp; window B:1,172K bp – 1,174K bp; *CFTR*:1,1711K bp – 1,1731K bp). They are both exclusively present in the top 1 % iHS of Balkars. Window A ranks 64th and window B 33th in the first 1% of the iHS excluding control population windows. Window A contains 21 SNPs, 10 of them having an iHS value above 2. Window B contains 20 SNPs, 8 of them having an iHS value above 2. Window A also contains *Cttnbp2*, window B also contains *ASZ1* (Tab. 6).

Table 5 For high altitude adaptation relevant enrichments detected in the top 1 % of Abkhazians and Balkars using DAVID gene enrichment. Analysis of the top 1 % of other Caucasus populations didn't yield any relevant enrichments.

population	enrichment	genes	p-value
Abkhazians	PPAR signalling pathway	*ACOX1*; *PLPT*; *RXRG*	0.027
Balkars	lung development	*ASZ1*; *Cttnb2*; *CFTR*; *MAPK8IP3*; *PROX1*	0.00029

Table 6 List of the two windows containing *CFTR*, present in the top 1 % of the Balkars. For each window additional information about significance and ranking is given (for explanation see Tab. 3).

position*	gene(s)	amount of SNPs	cut	mean	p-value	ranking**
chr.7 1,171K bp	*CFTR*; *ASZ1*	20	0.40	1.43	0.0055	64
chr.7 1,173K bp	*CFTR; CTTNBP2*	21	0.48	1.84	0.0026	33

* the given position x indicates the middle point of the window. The interval of the window is [x -100K bp; x + 100K bp]
** ranking indicates the position of the window in the ranking without excluding the windows that are also present in the control groups. Positions therefore can be > 100.

3.3 Candidate gene approach

The top 1 % of iHS windows in every Caucasus population excluding hits from control groups were screened with the individually selected gene lists (Supplementary Tab. 2 and Supplementary Tab. 3). Three genes in the top 1 % of the Abkhazians iHS hits could be identified. *Caveolin 1* (*Cav1*) is annotated with the GO term "hypoxia response". *RPA3* and *SMG6* are in the longevity list with the annotation "telomere maintenance". The screening of the top 1 % of the other populations didn't yield any results in both hypoxia and longevity related terms. The statistics of the windows containing the three aforementioned genes are listed in Table 7. The window with *Cav1* additionally contains *Caveolin 2* (*Cav2*). An overall summary of genes detected in Caucasus populations related to high altitude is given in Table 8.

Table 7 List of putatively relevant genes detected in top 1 % of Abkhazian data set via candidate gene list screening. Additional information about corresponding windows is provided.

	Cav1	***RPA3***	***SMG6***
GO term	hypoxia response	telomere maintenance	telomere maintenance
subsequent information refer to the corresponding window of the gene			
position*	chr. 7 1,161K bp	chr 7 7,700K bp	chr 17 2100K bp
additional gene	*Cav2*	*MIOS*	-
amount of SNPs	29	62	29
cut	0.38	0.24	0.35
mean	1.56	1.36	1.62
p-value	0.0085	0.0097	0.0111
ranking**	91	106	128

* the given position x indicates the middle point of the window. The interval of the window is [x -100K bp; x + 100K bp]
** ranking indicates the position of the window in the ranking without excluding the windows that are also present in the control groups. Positions therefore can be > 100.

Table 8 Summary of all genes that have been detected in Caucasus populations and are related to high altitude adaptation.

	selected gene
Abkhazians	
candidate gene approach	*Cav1*
PPAR signalling pathway*	*ACOX1*
	PLPT
	RXRG
Armenians	-
Balkars	
lung development*	*ASZ1*
	Cttnb2
	CFTR
	MAPK8IP3
	PROX1
Chechens	-
Georgians	-
Kumyks_Adyghei	-
Lezgins_NorthOssetians	-

* The left column indicates whether the gene has been detected by candidate gene approach or by gene enrichment.

3.4 Unique window for Caucasus populations

Only one window could be detected as being shared at top 5 % in all of the examined Caucasus populations while not being found in the top 5 % of the control groups (French and Lebanese_Syrians). It is located on chromosome 3 at position 11500 Mb. The ranking shows that it is especially significant in Armenians and Kumyks_Adyghei with a p-value < 0.003 (Tab. 9). One gene is present in the window: *PLCXD2* (Phosphatodylinositol-specific phospholipase C, X containing domain 2).

Table 9 List of the significance of the window unique in the first 5 % of the Caucasus populations. The window (position: Chr. 3; 11500 Mb) is not present in the top 5 % of the control groups.

	ranking	**p-value**	**cut**
Abkhazians	144	0.013788	0.238095
Armenia	27	0.002072	0.386364
Balkars	493	0.048337	0.152174
Chechens	59	0.005183	0.292683
Georgians	414	0.040852	0.166667
Kumyks_Adyghei	31	0.002481	0.350000
Lezgins_NorthOssetians	181	0.017575	0.227273
French	2053	0.20456	0.076923
Lebanese_Syrians	726	0.072733	0.157895

4 Discussion

4.1 Constraints in studying genetic adaptations

Before going into detail about possible individual genetic adaptations, several constraints will be discussed. First of all, the constraints in studying genetic adaptation in general will be considered. Secondly, constraints that originate from the study design have to be taken into consideration. In the end, the problems that are caused by iHS calculations will be analysed in more detail. All these constraints should be applied to every potential genetic adaptation that will be discussed subsequently.

4.1.1 General constraints in studying genetic adaptations

Omnipresent problems in evolutionary studies are demographic events and associated genetic drift. However, while demographic events and drift would be expected to affect all loci in the genome then the iHS method emphasizes particularly on those genetic regions where the derived allele has higher haplotype homozygosity than the ancestral allele. Nevertheless, while the common practice is to focus on the tails of the empirical distribution of iHS we do not actually know how much of the genome has been affected by selection. Even if no selection was involved we would be able to pick up the top 1% or top 5 % genes and conversely, if much of the genome was affected by selection, these criteria would be too conservative.

Even if there is a high certainty that an allele was selected, the next question is what kind of function this alteration results in. Functional studies and association studies can help here. However, pleiotropic effects (alteration in a gene can result in multiple different phenotypic traits) and epistatic effects (one gene masks, interferes with or enhances the expression of the other gene) complicate these studies (Bamshad and Wooding 2003).

In the end, even if phenotypic trait can be clearly assigned to one distinguished alteration, the nature of selection still remains subject to interpretation.

4.1.2 Constraints imposed by this study

One issue in studying high altitude adaptation in the Caucasus was already outlined in the introduction: have the populations lived in an altitude that was high enough to cause hypoxic adaptations? 2000 m is only a middle-hypoxic environment and most of high altitude adaptation studies were carried out with populations living at 3000 m or more. However, genetic high altitude adaptation of Daghestani population living at 2000 m has been shown previously (Pagani, Ayub et al. 2012). In general, one has to bear in mind that hypoxia and high altitude are gradual and a threshold defining a limit under which high altitude adaptation doesn't occur is only an approximation (see also Fig. 1).

The low-lying samplings points are a source of error. However, as already discussed in the introduction, several Caucasus populations have a long history of living in high altitude (~2000 m) and have undergone recent migration events. Nevertheless, there remains a bit of uncertainty about their actual origins.

Another problem is caused by Caucasus' position and history. Its position is central and its region has undergone several migration and immigration events. Even if isolation is high in mountainous villages it is incontestable that admixture occurred several times in Caucasus history. Therefore, gene flow was high which complicates the search for natural selection. In contrast, the Tibetan Plateau is one of the most isolated regions in the world and more suitable for detection of high altitude adaptation (Moore 2001).

4.1.3 Constraints of iHS calculations

iHS calculations have to encounter several constraints. First of all, it can only detect alleles that didn't reach fixation. A test which is somehow complementary to iHS is XP-EHH. This test doesn't use ancestral alleles to compare with but alleles from control populations. It detects sweeps that have high frequency or are fixed.
The next problem in evaluating iHS data is the fact, that significance is very difficult to measure. iHS values are transmitted into empirical p-values. However, iHS ranking is really variable. Important SNPs with a strong iHS value get lost, if they are not surrounded by SNPs that also have high values. This is because the p-value depends predominantly on the cut-value, which depends on the average of the iHS values of all the SNPs in a given window. All the significant SNPs that are by coincidence in windows with an amount of SNPs < 20 are also lost. In total, some information might be significant but get lost during analysis.
Thirdly, it is difficult to refer from a significant window, which genes actually are affected. Because iHS statistics are carried out for windows with a length of 200 kb, it is difficult to refer from these windows the selected haplotypes. If we want to find out about the actual selected haplotype, it would be important to increase the resolution.

4.2 High altitude adaptation

4.2.1 Do Abkhazians and Balkars have a high altitude residence background?

The populations in which putatively relevant genes could be detected are Abkhazians and Balkars (Tab. 8). Therefore it is important to discuss initially if it is theoretically possible that high altitude adaptation occurred. In other words, have they lived for a sufficient amount of time in an altitude that is high enough.

Abkhazians
Abkhazians' sampling point is 15 m high. However, Abkhazians belong to the aborigines of Caucasus and have deep roots in residence in the mountainous region of Abkhazia (ranging 3500 – 4000 m on the Eastern border) (Kolga, Tonurist et al. 2001). Its long residence and low admixture with other populations is also reflected in the ADMIXTURE plot, in which Abkhazians nearly do not share any ancestry with French, Han Chinese or Yoruba and only a few share some minor ancestry with Palestinians (Fig. 10).

Balkars
Balkars' sampling point is 715 m high. The origins of Balkars are not well understood, but can be traced back to Turkic tribes migrating to the North Caucasus by the end of the fourth century B.C. when they mixed with indigenous Caucasian peoples (Minahan 2002). However, their settlements since then have been in the high mountains, which is also reflected by their ancestral Russian name "Mountain Tatars" (Cole 2011). Only in 1944 under Stalin's rule, the Balkars were shipped east with a return to their homeland in 1956 (Minahan 2002). Nonetheless they were not allowed to settle in their former villages in the mountains, but in the foothills (Cole 2011). This would explain the low altitude sampling location. Nevertheless, even if the Balkars ancestors lived in high altitude, it is questionable if the short residence period was sufficient for high altitude adaptation to occur (~2 500 years). When compared to the other time ranges (Tibetans: 25 000 years; Andeans: 11 000 years (Beall 2007)), Balkars' time period seems indeed short. However, a microevolutionary study carried out in 4000 m of altitude has shown that Tibetan women with high oxygen saturation mutations had a higher offspring survival (Beall, Song et al. 2004). This indicates that adaptation is a process with

significant impact at the scale of a single generation and that advantageous mutations affecting offspring survival at high altitude may undergo rapid frequency change. This makes it plausible that Balkars in the time length of 2500 years also have encountered hypoxic selection pressure influencing the reproductive success and resulting in genetic adaptation. If the alleles that increase offspring survival at high altitude stages of population history are not maladaptive during the times when the population lives at lower altitude such allelic changes may persist, albeit being subject to relatively more stochastic changes due to drift and admixture than among populations continuously living at high altitude.
Hence, there are several uncertainties about origin and altitude living in Abkhazians and Balkars. However, several facts suggest that high altitude adaptation theoretically could explain the enrichment of hypoxia genes in the two populations in this study.

4.3 DAVID gene enrichment

4.3.1 Enrichment for the PPAR pathway in Abkhazians

The PPAR proteins (PPAR α, β, γ) are nuclear hormone-binding proteins that play an important role in lipid metabolism. PPARα interacts with components of the HIF pathway and can be related to high altitude adaptation in Tibetans. For instance, Simonson et al. (2010) reported that an alteration in *PPAR*α is significantly related with low Hb concentration in Tibetans highlanders. The gene shows significant signs of positive selection with an iHS value of 3,58 (Simonson, Yang et al. 2010).
PPARα itself is not detected in our case, however, three genes involved in the PPAR pathway are. The question arises how alteration in lipid metabolism can be related to hypoxia and high altitude. Correlation between hypoxia and adipogenesis has been detected in several studies: *PPARa* expression is inhibited by HIF1 during hypoxia in mice (Simonson, Yang et al. 2010). *PPAR*γ is likewise inhibited during hypoxia in mouse embryonic fibroblasts (Yun, Maecker et al. 2002). This correlation between hypoxia and adipogenesis can also be observed in high altitude training resulting in body fat reduction in humans. Under hypoxia, glycolysis is augmented to maintain energy homeostasis, while fatty acid oxidation is impeded (Yun, Maecker et al. 2002). Energy storage in form of adipogenesis therefore becomes less necessary. While this response might be advantageous for a direct response to hypoxia, under chronic hypoxia, this response might by harmful for the body. In seasonal climates, the ability to store excess energy is essential. Furthermore, high altitude is always related to low temperature, and fat tissue serves as a very important insulator.
One has to be aware that our methods (without any further experimental studies) can't refer the actual biological function from a selected haplotype. Therefore, the above-mentioned reasons are only assumptions. We can't tell if the alteration of the genes lead to a weakened hypoxia response in lipid metabolism. We furthermore cannot tell if the iHS genes are indeed enriched for the PPAR signalling pathway as the p-value (p-value = 0.03) is very high. If there is indeed a selected enrichment, additionally we don't know if this selection occurred in response to living in high altitude. Further investigations have to be carried out to solve some of these questions.

4.3.2 Enrichment for "lung development" in Balkars; the *CFTR* gene

The lung development, morphology and ventilation play an important role in high altitude adaptation. Andean highlanders have a larger lung volume after having grown up in high altitude environment compared to lowlanders who have grown up in high altitude (Brutsaert, Soria et al. 1999). It is therefore very likely that altered genes

correlated with lung development evolved in response to hypoxia. The possible meaning of this enrichment and especially the role of CFTR will be discussed in this section.

Significance of "lung development" enrichment

First of all, it has to be noted that the significance of the enrichment is very controversial. Although it has a p-value of 0.00029, it is evident that the enriched genes all have other important functions. *ASZ1*, for instance, is exclusively expressed in germ cells and is strongly related to infertility (Yan, Rajkovic et al. 2002). Even though *PROX1* is expressed in the embryonic lung, its main function is regarded to be the development of the CNS. Therefore, it can't be excluded that because of the pleiotropic effects the mentioned genes were selected for other reasons than for an altered lung development. This is especially true for genes involved with fertility. Fertility is subject to strong selective pressure and nearly in every enrichment of the different populations, genes were enriched for terms such as "sperm motility". These other enrichments however have not been analysed in further detail because they are not directly relevant for high altitude adaptation. It is also very striking that *ASZ1* and *Cttnbp2* are located next to *Cftr* on chromosome 7 in the same two windows. It is very likely that *Cttnbp2* for instance was only annotated with the term "lung development" because of its proximity to *CFTR*. Furthermore, it is very unlikely that two different genes in proximity were both target of selection. Therefore it is questionable if indeed SNPs of *ASZ1* and *Cttnbp2* are responsible for the high iHS values. It is much more likely that *CFTR* was target of selection and the cause for the high iHS values found in the two windows. In total, one can say that the discussed enrichment has weaknesses.

Even if the relevance of this special enrichment is questionable, the alteration of the *CFTR* gene is undoubtedly an interesting finding and the possible biological meaning was therefore investigated in more detail. Of course, one has to be aware of the fact, that there are three genes in the two significant windows and one has to do further investigations to really find out, which gene was target of positive selection.

CFTR and its role in hypoxia response

The *CFTR* gene is well known because of its connection with cystic fibrosis. The *CFTR* gene codes for an ion channel, which regulates the chloride transport in epithelial tissue and is crucial for the composition and liquidity of mucus. When the CFTR protein is degraded (as in the case of cystic fibrosis), the mucus becomes too viscous. In lungs, this results in an obstruction of the pulmonary glands and inhibition of mucociliary clearance and oxygen delivery (Guimbellot, Fortenberry et al. 2008). Degradation of the channel leads therefore to a worsening of oxygen delivery. Indeed, on a molecular basis, CFTR has a binding site for HIF-1 (Zheng, Kuhlicke et al. 2009). Surprisingly, this correlation is negative, hypoxic signals lead to repression of CFTR mRNA, protein and function (Guimbellot, Fortenberry et al. 2008). For instance, a 60 % loss of *CFTR* mRNA in mountaineers exposed to high altitude was demonstrated (Mairbäurl, Schwöbel et al. 2003). The cause of this kind of hypoxia response lies presumably in the energy homeostasis of the cell. Vectoral Cl^- -Ion transport through CFTR transport is an energy-dependent process and it is very likely that the transport is down regulated in order to conserve energy (Guimbellot, Fortenberry et al. 2008).

Here again, this response to hypoxia might be contra productive when hypoxia is chronic. The side effect of the down regulation of the expression of *CFTR* is best shown by the severe symptoms of cystic fibrosis. Also, in a study about HAPE, HAPE-susceptible subjects showed the discussed down regulation of *CFTR* and other ion channel expression while in the control group showing no HAPE symptoms this hypoxia

response was milder (Mairbäurl, Schwöbel et al. 2003). This suggests that the hypoxia response of *CFTR* expression alteration might be disadvantageous in a high altitude environment and the *CFTR* gene might have been under selection.
In total, one has to be aware that these are all considerations with a lot of uncertainties. If the *CFTR* gene has indeed been under natural selection, we still don't know what the biological function of this alteration is. We don't know if it affects the HIF regulation and if it does indeed, with which result. Of course, besides these uncertainties there are other general open questions discussed in the introduction of the discussion, which applies for every detected window or gene.

4.4 Candidate genes

4.4.1 *Cav1* and *Cav2* in Abkhazians

Cav1 and *Cav2* are both annotated with hypoxia response. It is again questionable if firstly, the two genes were target of selection and secondly, if they were indeed, they were selected in response to hypoxia response. Despite these uncertainties, it is worth to discuss the biological meaning and possible adaptation to hypoxia of the two genes.

Cav1 and its role for NO synthesis
Cav1 and *Cav2* are colocalized genes that code for two related proteins Cav1 and Cav2 that form hetero-oligomers and are the principal component of caveolae (Fra, Mastroianni et al. 1999). Caveolae are small invaginations of the plasma membrane and have several functions in signal transduction. In the context of hypoxia, Cav1 interacts with eNOS while eNOS is located in the caveolae (García-Cardeña, Fan et al. 1996). It is believed that while the two proteins Cav1 and eNOS are associated, eNOS is in an inactive state (Gratton, Fontana et al. 2000). Cav1 also regulates NOS2A. The interaction increases the degradation of NOS2A (Pautz, Art et al. 2010). In both cases, the interaction with Cav1 leads to a down regulation of nitric oxide (NO). NO is a signalling molecule with a lot of different physiological functions. It plays a crucial role in vasodilatation, relaxation of arterial smooth muscles and increased blood flow (Bigham, Bauchet et al. 2010). In several studies NOS, especially NOS2A could be related to hypoxia and high altitude. For instance, besides Cav1, it was found that HIF-1 is essential for the regulation of NOS2 gene transcription in pulmonary endothelium (Palmer, Semenza et al. 1998). NO production is increased in Tibetans resident at 4200 m compared to lowlanders (Erzurum, Ghosh et al. 2007). The most striking finding is that NOS2A exhibit strong signature of recent positive selection in Andeans (Bigham, Bauchet et al. 2010).

Cav1, NO and their physiological significance in hypoxia response

The question arises which are the possible reasons for the aforementioned correlations. Which role plays NO in the response to hypoxia? As NO causes vasodilatation, the blood circulation is improved. An improved blood circulation is especially important in pregnant women. As discussed in the introduction, a decisive factor in high altitude is the reduced foetal growth due to hypoxia. It could be shown that an increase in blood flow to the uteroplacental circulation in Andean and Tibetans is crucial in protecting them from a reduction in fetal growth (Bigham, Bauchet et al. 2010). I was also shown that a prolonged blockade of nitric oxide synthesis in gravid rats leads to an intrauterine growth retardation (Molnár, Sütö et al. 1994). To sum up, NO plays a crucial role in the blood supply for the foetus.

Besides the uterus blood supply, the synthesis of NO is also crucial in the pulmonary vessels. The severe high altitude disease HAPE is caused through pulmonary hypertension. For instance, vasodilators (amongst others NO) can be used to treat HAPE, reducing increases in pulmonary artery pressure and improve gas exchange (Bigham, Bauchet et al. 2010). The regulation of NOS2A by HIF could here be a natural physiological response to counteract the hypertension. The role of Cav1 in this case goes further than only regulating NOS. Cav1 is also involved in regulating K^+-channel function, which through different steps also acts on contraction of the smooth muscles (Murray, Insel et al. 2006). The correlation of Cav1 with lung hypertension is also shown by the fact that patients with idiopathic pulmonary hypertension have increased Cav1 mRNA and protein expression.

Collectively these facts suggest that Cav1 plays a crucial role in vasodilation and pulmonary hypertension. It down regulates NO synthesis and alteration in its expression leads to pulmonary hypertension. It is likely that an alteration in this gene evolved in response to high altitude. One could for instance assume that the alteration in the *Cav1* gene in Abkhazians leads to a less strong pulmonary hypertension. Furthermore alteration in the gene via NO synthesis could indeed improve the blood supply of the fetus. Traits directly affecting the reproductive success are always especially strongly selected. Despite these facts, one has to bear in mind, that no functional information about the SNPs is given. Therefore, any information about the possible role of *Cav1* in high altitude adaptation are only assumptions. All the biological correlation of *Cav1* with hypoxia response and HAPE justify however these assumptions. In particular, a very striking fact is that NOS2A, a protein downstream of Cav1 has already shown signatures of selection in another high altitude population, the Andeans.

In total, there are two findings for Abkhazians that can be related to high altitude adaptation. On the one hand, there is an enrichment for the PPAR pathway, on the other hand, *Cav1* and *Cav2* have signature of positive selection. Both findings are related to the HIF pathway. The interesting issue about both these findings is also that they are not directly related to pulmonary and haematological systems. Even if I mainly focused on the pulmonary and haematological impact, one also has to bear in mind that high altitude affects the whole body. Therefore, the whole body has to react. This is reflected in the PPAR pathway, as it affects the whole lipid metabolism. Without doubt the vasodilatation is very important in the lungs. However, as I discussed, the blood supply for the foetus is also crucial and therefore systemic changes in response to high altitude are also important.

4.5 Longevity

4.5.1 Genetic evidence for longevity in Abkhazians

Regardless of the controversies surrounding the longevity evidence in the Caucasus two longevity genes were detected in Abkhazians having signatures of positive selection. Therefore, the possible biological function of the two genes, *RPA3* and *SMG6* is provided below.

RPA3

RPA3 codes for the protein RPA3, a subunit of RPA. Together with RPA1 and RPA2, these subunits form the heterotrimeric replication protein A (RPA). In our case it is important to note that RPA3 is crucial for RPA's function. An alteration in RPA3 therefore affects the whole functionality of RPA. RPA is a single strand DNA binding protein and plays an essential role in replication, recombination, DNA repair, and telomere maintenance (Salas, Petruseva et al. 2009). For instance, RPA is involved in nucleotide excision repair and is also required for mismatch repair *in vitro* (Nehlin, Skovgaard et al. 2000). DNA repair mechanisms play a key role in ageing (Lombard, Chua et al. 2005). This is best demonstrated by the fact that most of the premature ageing syndromes (such as Werner syndrome, Cockayne's syndrome and xeroderma pigmentosum) are caused by defective DNA repair. Striking evidence for ageing associtation of RPA is the interaction of RPA with the Werner syndrome protein (WRNp). This protein is a DNA-helicase and when non functional, the gene causes the Werner syndrome (discussed in the introduction). RPA helps WRNp to unwind DNA through direct protein-protein interaction (Nehlin, Skovgaard et al. 2000). This facilitates for instance DNA-replication. In summary, RPA is an important contributor for DNA repair and replication and can therefore be closely linked to ageing.

SMG6

The *SMG6* gene codes for the Telomerase-binding protein EST1A. EST1A is one of two EST1 proteins found to be associated with telomerase. An increase in EST1A expression leads to a significant reduction in telomere length (Lundblad 2003). Furthermore, overexpression leads also to an uncapping of the chromosomes, suggesting that EST1A contributes to chromosome end protection (Lundblad 2003). Reduced telomere length is associated with a reduced life expectancy. For instance, average telomeres lengths in fibroblasts from Hutchinson-Gilford Syndrome patients were found to be shorter than the one from controls (Decker, Chavez et al. 2009).

4.5.2 Possible explanations for the alteration in *RPA3* and *SMG6*

The question arises which consequences could have the given alteration in RPA3. First of all it is important to note that it is a very conservative protein in eukaryotes (Salas, Petruseva et al. 2009). In general, mutations causing a deteriorated functionality of the protein are more likely to occur. But if the Abkhazians indeed had a higher life expectation the alteration must have other consequences. It could for instance lead to an improved functionality of RPA3, which would lead to an improved DNA damage repair and telomere maintenance, leading to a retarded ageing. The same question can be applied to the alteration of *SMG6*. Here an overexpression leads to shortened telomere length. This suggests a negative regulation through EST1A. When we want to explain an alteration in terms of elevated life expectancy, one could for instance imagine that the selected haplotype has a reduced affinity to telomerase.

Cancer avoidance is another possible explanation of how the selected genes could contribute to longevity. Terms like replication, telomeres, telomerase activity and senescence are closely linked to the development of cancer. It could also be that the alterations of the genes in questions cause a better avoidance of cancer with for example a faster entering in the replicative senescence. If there is a stronger restriction in terms of replication and immortality, cancer is less likely to develop and to spread. Therefore longevity could be explained by the fact that people with the selected genes are less likely to get cancer. Especially when one considers that cancer is an age-related disease and advanced age is associated with a higher cancer risk (Pavlidis, Stanta et al.). Indeed, a new study investigating centenarians and cancer showed that in centenarians cancer is a relatively uncommon disease and a rare cause of death (Pavlidis, Stanta et al.).

4.5.3 Explanations for longevity in high altitude environment

So far, possible genetic evidence that could cause longevity was discussed. In this section, present different possible explanations for longevity in high altitude environment will be considered. The first one is an actual explanation for selection pressure towards a longer life. The second one assumes that longevity is due to environmental factors excluding any genetic determinants. The last explanation considers longevity as a side-effect of the adaptation to hypoxia.

Longevity as a selected trait

If there are indeed selected genes that cause the body to live longer, the question arises why these genes were selected in Abkhazians. It is possible that in the harsh environment of high altitude, a different life history strategy was advantageous. High altitude environment is linked to nutritional constraints and severe climate changes. These constraints are normally linked with a slower growth and later reproductive age. However, a slower growth and a late reproductive success is always linked with the risk to die before reproduction. There is a balance between growing as slow as possible and reproducing as quick as possible. It could be that the balance is shifted because the risk of dying is decreased in an isolated high altitude environment. First of all, infectious diseases are less likely to occur in isolated small populations. The population size is not big enough to keep a pathogen endemic. Furthermore, in the cold climate there are less opportunities/hosts for parasites to grow and to spread. The only constraint would be in this case their own body and the ageing process. If it is advantageous to grow more slowly and the body finds a way to delay ageing symptoms and cancer, the life trajectory can shift towards a longer life resulting in a higher reproductive success. It has to be noted, that this line of argument doesn't have any evidence. There is, for instance a lack of life history trait studies in high altitude population, therefore the thesis can't be substantiated by any concrete data.

Longevity as a consequence of environmental factors

In this argument, it was also assumed that in an isolated high altitude environment there is a lower risk of getting diseases and dying early. These arguments can of course be utilized for another explanation, excluding any genetic determinants. If there is indeed a risk reduction in dying early, this could be the explanation for a high amount of elderly people in these populations, without any selection pressure for longevity and without any selected genes that can be related to longevity. In this case, there must be other origins for the alteration of *SMG6* and *RPA3*of than the selection for longevity.

Longevity as a consequence of adaptation to hypoxia

A third explanation for longevity in high altitude becomes possible when we consider the possible impacts of adaptation to hypoxia. The longevity could only be a side effect of the adaptation to high-altitude. This means that throughout the adaptation to high altitude, the people started to live longer. A possible explanation would be that for the purpose of saving oxygen the metabolism in general changed. Let's assume that adaptation to hypoxia not only consists of improving the oxygen supply but also of reducing the general oxygen need. This would imply a decrease in metabolism. As a consequence, reactive oxygen species (ROS), a normal by-product of metabolism, would be reduced. ROS can cause several types of DNA damages and plays a key role in ageing (Lombard, Chua et al. 2005). With a reduction of oxidative stress, the organism would automatically increase its life expectancy. This approach is underpinned by the fact that there is a general trend that organisms with a high altitude habitat live longer. For instance, within certain floral species (Ranunculaceae, forb species) the longevity of high altitude populations is higher compared to low altitude populations (Zhang et al 2006) (von Arx, Edwards et al. 2006). This phenomenon could also be observed in the animal kingdom. Marbled newts from high altitude reach older age than animal of the same species in low altitude (Caetano and Castanet 1993). However, this general trend could also be explained by a convergent adaptation to high altitude. It could be that the harsh conditions of high altitude (poor nutrition, hypoxia, low climate) provoke in different organisms and species similar changes in life history traits. This would favour the assumption that indeed the detected changes in the DNA repair genes were positively selected. But nevertheless, it remains yet to be shown that high altitude populations show any tendency to live longer.

4.6 Unique selection signal in the Caucasus populations

One window was detected in the first 5 % of all the Caucasus populations but not in the control groups. The corresponding gene is *PLCXD2*, a Phospholipase C enzyme whose specific function has yet to be revealed. It is only known to interact with phosphatidylinositol and might therefore play a role in signal transduction pathways. Further studies have to be carried out to infer possible meanings for the selection of this gene.

The Caucasus populations have a complex population structure with diverse backgrounds (also reflected in PCA and Admixture). Therefore, this approach is in general a difficult one and also not suited to detect high altitude adaptation, as Adyghei and Kumyks are lowlanders. The aim was to detect selected genes that were unique to Caucasus populations as a regional cluster of ethnic groups living in the same geographic area. The fact that window's significance is highest in Armenians and Adyghei and Kumyks is surprising because they all have different migration backgrounds. Armenians are known to have lived there for over 4000 years. Their roots might lay deep in human ancestry, as there is archaeological evidence that Armenian highlands is one of the earliest site of civilization (~4000 BC) (Samuelian, 2000). Adyghei are also believed to be long resident in the North Caucasus but information about their origin are limited in literature. In opposite, Kumyks are one of the youngest ethnic group (600 – 700 years) in the Caucasus with a Turkic-speaking background (Bulayeva, Marchani et al. 2008). However, it is believed that Kumyks might be partly of Daghestanian local origin (Matras, McMahon et al. 2006). In summary, it is very difficult to draw any conclusions, especially because Kumyks and Adyghei's iHS values were calculated together, but have different population history. This is because the main

study design was created for high altitude analysis, and therefore the elevation had priority.
A different study design, in which population structure would stronger be taken into account, would have the potential to reveal more genes that were exclusively selected in ancestral Caucasus natives. For instance, one could analyse what is unique for Armenians, Georgians and Abkhazians as they all cluster closely together in PCA plot (Fig. 9). One could for example investigate in detail the immune system. Both high altitude living and insulation has impact on the nature of pathogens that humans encountered. The cold climate decreases the amount of parasites, the small populations size and insulation decreases has influence on the pathogenicity of bacteria and viruses. However this analysis is beyond the scope of this study.

5 Outlook

In summary, Abkhazians and Balkars show signatures of selection for genes that are likely to be related to high altitude adaptation. Additionally, two genes involved in DNA replication and telomerase activity could be detected in Abkhazians. These findings are however subject to uncertainties and require some further examination to make it significant.

New methods improving the detection of selection

There are several new approaches that are likely to improve this type of studies in future. Firstly, a problem that was also encountered in this study is the admixture of different ethnics. The detection of signals for high altitude adaptation could highly be improved by only taking these bits of haplotypes that originates from ethnics adapted to high altitude. A similar approach was made in an admixture study investigating Mexican individuals with African, European, and American indigenous backgrounds (Johnson, Coram et al. 2011). They were able to infer ancestral components of these admixed genes and could trace back the demographic history of their African, European and American ancestors. These "virtual genomes" could in future be used to encounter the problem of admixture. This is especially important in a world where more and more admixture takes place.

In this study, iHS was used to detect signatures of selection. The problem here is to refer the selected haplotype(s) from the broad region of selection indicated by the calculations. As long as only regions of selection are detected, possible impacts on the phenotype remain uncertain. A combined approach using several different methods has been shown to highly increase the resolution of these tests (Grossman, Shylakhter et al. 2010). In future, it will therefore be possible to approach more and more the exact location of selection, especially with the background of more efficient and cost-effective sequencing methods.

On which elements selection acts on

If we want to continue to find out more about selection and human evolution we have to go back to the very beginning and ask the question on which levels the selection acts on. There are two requirements for an element to be selected: it has to be heritable and is musts somehow influence the phenotype. It is likely that current studies put in general too much emphasis on genes themselves. It is important to note that only a tiny percentage of the genome is genic DNA. Amino acid changes in proteins can't explain the individuality of every person on the world. There is a myriad of other things that make up the biological individuality of a person. The first things evidently are mutations in cis-regulatory elements. As correlation between regions in the genome and the linked actual gene is mostly unknown, important regulatory haplotypes are very likely to be overlooked.

A widely controversial topic in this context is also the so-called "junk" DNA. Repetitive elements, such as SINEs, LINEs, and microsatellites, used to identify individuals might indeed affect one's individual phenotype. If they indeed influence for instance the regulation of proteins and therefore the phenotype they are also subject to selection. If we want to understand the differences between populations, we therefore might also carry out studies investigating theses parts of the genome. Here, there is also a problem that is imposed by our currently used next-generation sequencing methods. The short reads (30 – 200 k bp in lengths) cause problems in putting together the true genome of an individual, with all its individual differences in repetitive sequences and copy number

variations. However, it could also be that these elements are indeed not relevant for the phenotype.
The last element that might be important for selection is epigenetic changes. It is still very unclear to which extent epigenetic changes are heritable. Epigenetic changes have the ability to severely change the expression pattern of genes and can therefore also have an impact on the phenotype (and therefore also be under selection). If heritability is indeed high, epigenetics would be a new level on which signs of selection could be detected.
It is undoubtedly, that in the end, the proteins and protein expression make up the phenotype. Therefore, an approach to overcome these other (so far overlooked) elements could be to emphasize protein expression pattern research. If we can link expression patterns of a cell with its whole genome (and not only genes) and epigenetic information, links between these two levels (genome and proteins) become more clear. In terms of high altitude, it is very likely to reveal differences by studying protein expression pattern in high altitude adaptation populations in more detail. Afterwards, one could trace backwards these differenced to potential selected haplotypes.

6 Bibilography

Adler, D. S., O. Bar-Yosef, et al. (2008). "Dating the demise: Neandertal extinction and the establishment of modern humans in the southern Caucasus." Journal of Human Evolution **55**(5): 817-833.

Bamshad, M. and S. P. Wooding (2003). "Signatures of natural selection in the human genome." Nat Rev Genet **4**(2): 99-111.

Beall, C. M. (2007). "Detecting natural selection in high-altitude human populations." Respiratory Physiology & Neurobiology **158**(2,Äì3): 161-171.

Beall, C. M. (2007). "Two routes to functional adaptation: Tibetan and Andean high-altitude natives." Proceedings of the National Academy of Sciences **104**(Suppl 1): 8655-8660.

Beall, C. M., K. Song, et al. (2004). "Higher offspring survival among Tibetan women with high oxygen saturation genotypes residing at 4,000 m." Proceedings of the National Academy of Sciences of the United States of America **101**(39): 14300-14304.

Bennett, N. G. and L. K. Garson (1986). "Extraordinary Longevity in the Soviet Union:Fact or Artifact?" The Gerontologist **26**(4): 358-361.

Bigham, A., M. Bauchet, et al. (2010). "Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data." PLoS Genet **6**(9): e1001116.

Bigham, A. W., X. Mao, et al. (2009). "Identifying positive selection candidate loci for high-altitude adaptation in Andean populations." Human genomics **4**(2): 79-90.

Browning, S. R. and B. L. Browning (2007). "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering." The American Journal of Human Genetics **81**(5): 1084-1097.

Browning, S. R. and B. L. Browning (2011). "Haplotype phasing: existing methods and new developments." Nat Rev Genet **12**(10): 703-714.

Brutsaert, T. D., R. Soria, et al. (1999). "Effect of developmental and ancestral high altitude exposure on chest morphology and pulmonary function in Andean and European/North American natives." American Journal of Human Biology **11**(3): 383-395.

Bulayeva, K., E. Marchani, et al. (2008). "Genetic bottleneck among daghestan highlanders migrating to lowlands." Central European Journal of Medicine **3**(4): 396-405.

Caciagli, L., K. Bulayeva, et al. (2009). "The key role of patrilineal inheritance in shaping the genetic variation of Dagestan highlanders." J Hum Genet **54**(12): 689-694.

Caetano, M. H. and J. Castanet (1993). "Variability and microevolutionary patterns in Triturus marmoratus from Portugal: age, size, longevity and individual growth." Amphibia-Reptilia **14**: 117 - 129.

Cheviron, Z. A. and R. T. Brumfield (2012). "Genomic insights into adaptation to high-altitude environments." Heredity **108**(4): 354-361.

Christensen, K. and J. W. Vaupel (1996). "Determinants of longevity: genetic, environmental and medical factors." Journal of Internal Medicine **240**(6): 333-341.

Colarusso, J. (1995). "Abkhazia." Central Asian Survey **14**(1): 75-96.

Cole, J. (2011). Ethnic Groups of Europe: An Encyclopedia, ABC-CLIO.

Coppedè, F. (2012). Premature Aging Syndrome
Neurodegenerative Diseases. S. I. Ahmad, Springer US. **724:** 317-331.

Decker, M. L., E. Chavez, et al. (2009). "Telomere length in Hutchinson-Gilford Progeria Syndrome." Mechanisms of Ageing and Development **130**(6): 377-383.

Erzurum, S. C., S. Ghosh, et al. (2007). "Higher blood flow and circulating NO products offset high-altitude hypoxia among Tibetans." Proceedings of the National Academy of Sciences **104**(45): 17593-17598.

Fra, A. M., N. Mastroianni, et al. (1999). "Human Caveolin-1 and Caveolin-2 Are Closely Linked Genes Colocalized with WI-5336 in a Region of 7q31 Frequently Deleted in Tumors." Genomics **56**(3): 355-356.

Fry, A. E., C. J. Trafford, et al. (2006). "Haplotype Homozygosity and Derived Alleles in the Human Genome." American journal of human genetics **78**(6): 1053-1059.

García-Cardeña, G., R. Fan, et al. (1996). "Endothelial Nitric Oxide Synthase Is Regulated by Tyrosine Phosphorylation and Interacts with Caveolin-1." Journal of Biological Chemistry **271**(44): 27237-27240.

Gratton, J.-P., J. Fontana, et al. (2000). "Reconstitution of an Endothelial Nitric-oxide Synthase (eNOS), hsp90, and Caveolin-1 Complex in Vitro." Journal of Biological Chemistry **275**(29): 22268-22272.

Grossman, S. R., I. Shylakhter, et al. (2010). "A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection." Science **327**(5967): 883-886.

Guimbellot, J. S., J. A. Fortenberry, et al. (2008). "Role of Oxygen Availability in CFTR Expression and Function." American Journal of Respiratory Cell and Molecular Biology **39**(5): 514-521.

Heidinger, B. J., J. D. Blount, et al. (2012). "Telomere length in early life predicts lifespan." Proceedings of the National Academy of Sciences **109**(5): 1743-1748.

Hosack, D. A., G. Dennis, et al. (2003). "Identifying biological themes within lists of genes with EASE." Genome Biology **4**(70).

Huang, D. W., B. T. Sherman, et al. (2008). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat. Protocols **4**(1): 44-57.

Huang, D. W., B. T. Sherman, et al. (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Research **37**(1): 1-13.

Jobling, M. A., M. Hurles, et al. (2004). Human Evolutionary Genetics: Origins, Peoples & Disease, Garland Science.

Johnson, N. A., M. A. Coram, et al. (2011). "Ancestral Components of Admixed Genomes in a Mexican Cohort." PLoS Genet **7**(12): e1002410.

Kenyon, C., J. Chang, et al. (1993). "A C. elegans mutant that lives twice as long as wild type." Nature **366**(6454): 461-464.

Kolga, M., I. Tonurist, et al. (2001). The Red Book of the Peoples of the Russian Empire. Talinn.

Kyucharyants, V. (1974). "Will the Human Life-Span Reach One Hundred?" The Gerontologist **14**(5 Part 1): 377-380.

Lacombe, C., N. Arous, et al. (1987). "A new case of HB Dagestan [alpha 60(E9)Lys–Glu]." Hemoglobin **11**(1): 39 - 41.

Lelashvili, N. G. and S. M. Dalakishvili (1984). "Genetic study of high longevity index populations." Mechanisms of Ageing and Development **28**(2,Äì3): 261-271.

Li, J. Z., D. M. Absher, et al. (2008). "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation." Science **319**(5866): 1100-1104.

Lombard, D. B., K. F. Chua, et al. (2005). "DNA Repair, Genome Stability, and Aging." Cell **120**(4): 497-512.

Lundblad, V. (2003). "Telomere Replication: An Est Fest." Current Biology **13**(11): R439-R441.

Mairbäurl, H., F. Schwöbel, et al. (2003). "Altered ion transporter expression in bronchial epithelium in mountaineers with high-altitude pulmonary edema." Journal of Applied Physiology **95**(5): 1843-1850.

Matras, Y., A. M. S. McMahon, et al. (2006). Linguistic Areas: Convergence In Historical and Typological Perspective, Palgrave Macmillan.

Migliaccio, E., M. Giorgio, et al. (1999). "The p66shc adaptor protein controls oxidative stress response and life span in mammals." Nature **402**(6759): 309-313.

Minahan, J. (2002). Encyclopedia of the Stateless Nations: A-C, Greenwood Press.

Molnár, M., T. Sütö, et al. (1994). "Prolonged blockade of nitric oxide synthesis in gravid rats produces sustained hypertension, proteinuria, thrombocytopenia, and intrauterine growth retardation." American journal of obstetrics and gynecology **170**(5): 1458-1466.

Monge, C. and F. Leon-Velarde (1991). "Physiological adaptation to high altitude: oxygen transport in mammals and birds." Physiological Reviews **71**(4): 1135-1172.

Moore, L. G. (2001). "Human Genetic Adaptation to High Altitude." High Altitude Medicine & Biology **2**(2): 257 - 279.

Moore, L. G. (2003). "Fetal growth restriction and maternal oxygen transport during high altitude pregnancy." High Altitude Medicine & Biology **4**(2): 141 - 156.

Murray, F., P. A. Insel, et al. (2006). "Role of O2-sensitive K+ and Ca2+ channels in the regulation of the pulmonary circulation: Potential role of caveolae and implications for high altitude pulmonary edema." Respiratory Physiology & Neurobiology **151**(2,Äì3): 192-208.

Nehlin, J. O., G. L. Skovgaard, et al. (2000). "The Werner Syndrome: A Model for the Study of Human Aging." Annals of the New York Academy of Sciences **908**(1): 167-179.

Pagani, L., Q. Ayub, et al. (2012). "High altitude adaptation in Daghestani populations from the Caucasus." Human Genetics **131**(3): 423-433.

Palmer, L. A., G. L. Semenza, et al. (1998). "Hypoxia induces type II NOS gene expression in pulmonary artery endothelial cells via HIF-1." American Journal of Physiology - Lung Cellular and Molecular Physiology **274**(2): L212-L219.

Palmore, E. (1984). "Longevity in Abkhazia: a reevaluation." Gerontologist **24**.

Pautz, A., J. Art, et al. (2010). "Regulation of the expression of inducible nitric oxide synthase." Nitric Oxide **23**(2): 75-93.

Pavlidis, N., G. Stanta, et al. "Cancer prevalence and mortality in centenarians: A systematic review." Critical Reviews in Oncology/Hematology(0).

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

Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet **38**(8): 904-909.

Purcell, S., B. Neale, et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." The American Journal of Human Genetics **81**(3): 559-575.

Salas, T. R., I. Petruseva, et al. (2009). "Evidence for direct contact between the RPA3 subunit of the human replication protein A and single-stranded DNA." Nucleic Acids Research **37**(1): 38-46.

Sampedro Camarena, F., G. Cano Serral, et al. (2007). "Telomerase and telomere dynamics in ageing and cancer: current status and future directions." Clinical and Translational Oncology **9**(3): 145-154.

Scheinfeldt, L., S. Soi, et al. (2012). "Genetic adaptation to high altitude in the Ethiopian highlands." Genome Biology **13**(1): R1.

Schoenmaker, M., A. J. M. de Craen, et al. (2005). "Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study." Eur J Hum Genet **14**(1): 79-84.

Simonson, T. S., Y. Yang, et al. (2010). "Genetic Evidence for High-Altitude Adaptation in Tibet." Science **329**(5987): 72-75.

Voight, B. F., S. Kudaravalli, et al. (2006). "A Map of Recent Positive Selection in the Human Genome." PLoS Biol **4**(3): e72.

von Arx, G., P. J. Edwards, et al. (2006). "EVIDENCE FOR LIFE HISTORY CHANGES IN HIGH-ALTITUDE POPULATIONS OF THREE PERENNIAL FORBS." Ecology **87**(3): 665-674.

Yan, W., A. Rajkovic, et al. (2002). "Identification of Gasz, an Evolutionarily Conserved Gene Expressed Exclusively in Germ Cells and Encoding a Protein with Four Ankyrin Repeats, a Sterile-Œ± Motif, and a Basic Leucine Zipper." Molecular Endocrinology **16**(6): 1168-1184.

Yi, X., Y. Liang, et al. (2010). "Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude." Science **329**(5987): 75-78.

Yun, Z., H. L. Maecker, et al. (2002). "Inhibition of PPARŒ≥2 Gene Expression by the HIF-1-Regulated Gene DEC1/Stra13: A Mechanism for Regulation of Adipogenesis by Hypoxia." Developmental Cell **2**(3): 331-341.

Yunusbayev, B., M. Metspalu, et al. (2011). "The Caucasus as an asymmetric semipermeable barrier to ancient human migrations." Molecular Biology and Evolution.

Zheng, W., J. Kuhlicke, et al. (2009). "Hypoxia inducible factor-1 (HIF-1)-mediated repression of cystic fibrosis transmembrane conductance regulator (CFTR) in the intestinal epithelium." The FASEB Journal **23**(1): 204-213.

7 Appendix

7.1 Gene list of top 1 % of the Caucasus populations

Supplementary Table 1 List of the genes in the top 1 % of the Caucasus populations, excluding the ones of the first 5 % present in the control groups.

Abkhazians		
Entrez ID	Official Gene Name	Gene Name
23033	dopey1	dopey family member 1
63027	SLC22A23	solute carrier family 22, member 23
22871	nlgn1	neuroligin 1
5894	RAF1	v-raf-1 murine leukemia viral oncogene homolog 1
7125	TNNC2	troponin C type 2 (fast)
3338	dnajc4	DnaJ (Hsp40) homolog, subfamily C, member 4
63925	ZNF335	zinc finger protein 335
5991	rfx3	regulatory factor X, 3 (influences HLA class II expression)
55287	TMEM40	transmembrane protein 40
84304	nudt22	nudix (nucleoside diphosphate linked moiety X)-type motif 22
23558	WBP2	WW domain binding protein 2
374897	SBSN	suprabasin
389362	Psmg4	proteasome (prosome, macropain) assembly chaperone 4
440699	Lrrc52	leucine rich repeat containing 52
8464	Supt3h	suppressor of Ty 3 homolog (S. cerevisiae)
257144	GCET2	germinal center expressed transcript 2
6119	Rpa3	replication protein A3, 14kDa
140686	Wfdc3	WAP four-disulfide core domain 3
90204	ZSWIM1	zinc finger, SWIM-type containing 1
90203	SNX21	sorting nexin family member 21
79750	ZNF385D	zinc finger protein 385D
1340	COX6B1	cytochrome c oxidase subunit Vib polypeptide 1 (ubiquitous)
10963	stip1	stress-induced-phosphoprotein 1
285335	SLC9A10	solute carrier family 9, member 10
63935	PCIF1	PDX1 C-terminal inhibiting factor 1
54468	MIOS	missing oocyte, meiosis regulator, homolog (Drosophila)
5360	Pltp	phospholipid transfer protein
6258	RXRG	retinoid X receptor, gamma
91107	TRIM47	tripartite motif-containing 47
140825	NEURL2	neuralized homolog 2 (Drosophila)
10005	ACOT8	acyl-CoA thioesterase 8
55327	LIN7C	lin-7 homolog C (C. elegans)
64978	Mrpl38	mitochondrial ribosomal protein L38
9215	LARGE	like-glycosyltransferase
55366	LGR4	leucine-rich repeat-containing G protein-coupled receptor 4
201292	Trim65	tripartite motif-containing 65
10430	Tmem147	transmembrane protein 147
201294	UNC13D	unc-13 homolog D (C. elegans)
23609	mkrn2	makorin ring finger protein 2
93099	Dmkn	dermokine
83706	fermt3	fermitin family homolog 3 (Drosophila)
2116	ETV2	ets variant 2
83707	Trpt1	tRNA phosphotransferase 1
11045	UPK1A	uroplakin 1A
3801	Kifc3	kinesin family member C3
79669	C3orf52	chromosome 3 open reading frame 52
347733	tubb2b	tubulin, beta 2B
1124	Chn2	chimerin (chimaerin) 2
27145	FILIP1	filamin A interacting protein 1
313	Aoah	acyloxyacyl hydrolase (neutrophil)
100134934	C17orf106	hypothetical protein LOC100134934
23354	HAUS5	HAUS augmin-like complex, subunit 5
140831	ZSWIM3	zinc finger, SWIM-type containing 3
28992	macrod1	MACRO domain containing 1
23293	SMG6	Smg-6 homolog, nonsense mediated mRNA decay factor (C. elegans)
116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1
3516	Rbpj	recombination signal binding protein for immunoglobulin kappa J region

85302	fbf1	Fas (TNFRSF6) binding factor 1
83698	CALN1	calneuron 1
54504	cpvl	carboxypeptidase, vitellogenic-like
79660	PPP1R3B	protein phosphatase 1, regulatory (inhibitor) subunit 3B
9378	NRXN1	neurexin 1
128497	C20orf165	chromosome 20 open reading frame 165
11065	UBE2C	ubiquitin-conjugating enzyme E2C
79171	rbm42	RNA binding motif protein 42
344805	TMPRSS7	transmembrane protease, serine 7
858	CAV2	caveolin 2
857	cav1	caveolin 1, caveolae protein, 22kDa
1018	Cdk3	cyclin-dependent kinase 3
495	ATP4A	ATPase, H+/K+ exchanging, alpha polypeptide
5476	CTSA	cathepsin A
5998	rgs3	regulator of G-protein signaling 3
51	ACOX1	acyl-Coenzyme A oxidase 1, palmitoyl
85451	unk	unkempt homolog (Drosophila)
1258	CNGB1	cyclic nucleotide gated channel beta 1
26330	GAPDHS	glyceraldehyde-3-phosphate dehydrogenase, spermatogenic
90025	Ube2cbp	ubiquitin-conjugating enzyme E2C binding protein
23769	FLRT1	fibronectin leucine rich transmembrane protein 1
Kumyks_Adygei		
Entrez ID	Official Gene Name	Gene Name
57630	SH3RF1	SH3 domain containing ring finger 1
81501	TM7SF4	transmembrane 7 superfamily member 4
100271229	RPL9P33	ribosomal protein L9 pseudogene 33
2898	GRIK2	glutamate receptor, ionotropic, kainate 2
7103	TSPAN8	tetraspanin 8
6095	RORA	RAR-related orphan receptor A
63892	THADA	thyroid adenoma associated
30815	ST6GALNAC6	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 6
57144	PAK7	p21 protein (Cdc42/Rac)-activated kinase 7
27090	ST6GALNAC4	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 4
138429	PIP5KL1	phosphatidylinositol-4-phosphate 5-kinase-like 1
9771	RAPGEF5	Rap guanine nucleotide exchange factor (GEF) 5
50631	YBX1P1	Y box binding protein 1 pseudogene 1
5873	RAB27A	RAB27A, member RAS oncogene family
23303	KIF13B	kinesin family member 13B
5756	TWF1	twinfilin, actin-binding protein, homolog 1 (Drosophila)
100270907	RPL6P12	ribosomal protein L6 pseudogene 12
79652	TMEM204	transmembrane protein 204
27145	FILIP1	filamin A interacting protein 1
89858	SIGLEC12	sialic acid binding Ig-like lectin 12
100049587	SIGLEC14	sialic acid binding Ig-like lectin 14
238	ALK	anaplastic lymphoma receptor tyrosine kinase
5629	PROX1	prospero homeobox 1
406887	MIRLET7E	microRNA let-7e
9046	DOK2	docking protein 2, 56kDa
946	SIGLEC6	sialic acid binding Ig-like lectin 6
8778	SIGLEC5	sialic acid binding Ig-like lectin 5
130271	PLEKHH2	pleckstrin homology domain containing, family H (with MyTH4 domain) member 2
28965	SLC27A6	solute carrier family 27 (fatty acid transporter), member 6
79745	CLIP4	CAP-GLY domain containing linker protein family, member 4
23233	EXOC6B	exocyst complex component 6B
54556	ING3	inhibitor of growth family, member 3
2066	ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)
4750	NEK1	NIMA (never in mitosis gene a)-related kinase 1
1807	DPYS	dihydropyrimidinase
9699	RIMS2	regulating synaptic membrane exocytosis 2
57585	CRAMP1L	Crm, cramped-like (Drosophila)
254827	NAALADL2	N-acetylated alpha-linked acidic dipeptidase-like 2
7728	ZNF175	zinc finger protein 175
407056	MIR99B	microRNA 99b
64478	CSMD1	CUB and Sushi multiple domains 1
79974	C7orf58	chromosome 7 open reading frame 58
399665	FAM102A	family with sequence similarity 102, member A
406910	MIR125A	microRNA 125a

387921	NHLRC3	NHL repeat containing 3
9355	LHX2	LIM homeobox 2
51187	RSL24D1	ribosomal L24 domain containing 1; similar to ribosomal protein L24-like
284288	RSL24D1	ribosomal L24 domain containing 1; similar to ribosomal protein L24-like
2201	FBN2	fibrillin 2
9742	IFT140	intraflagellar transport 140 homolog (Chlamydomonas)
1305	COL13A1	collagen, type XIII, alpha 1
203	AK1	adenylate kinase 1
57706	DENND1A	DENN/MADD domain containing 1A
90861	HN1L	hematological and neurological expressed 1-like
1846	DUSP4	dual specificity phosphatase 4
84216	TMEM117	transmembrane protein 117
80031	SEMA6D	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6D
23162	MAPK8IP3	mitogen-activated protein kinase 8 interacting protein 3
8818	DPM2	dolichyl-phosphate mannosyltransferase polypeptide 2, regulatory subunit
23039	XPO7	exportin 7
2022	ENG	endoglin
100132336	GFRA2	similar to GDNF family receptor alpha 2; GDNF family receptor alpha 2
2675	GFRA2	similar to GDNF family receptor alpha 2; GDNF family receptor alpha 2
Armenians		
Entrez ID	Official Gene Name	Gene Name
5873	Rab27a	RAB27A, member RAS oncogene family
1602	Dach1	dachshund homolog 1 (Drosophila)
54808	DYM	dymeclin
284254	C18orf26	chromosome 18 open reading frame 26
254827	Naaladl2	N-acetylated alpha-linked acidic dipeptidase-like 2
3680	ITGA9	integrin, alpha 9
57706	dennd1a	DENN/MADD domain containing 1A
5019	oxct1	3-oxoacid CoA transferase 1
3751	KCND2	potassium voltage-gated channel, Shal-related subfamily, member 2
10017	Bcl2l10	BCL2-like 10 (apoptosis facilitator)
25827	Fbxl2	F-box and leucine-rich repeat protein 2
55274	Phf10	PHD finger protein 10
6991	TCTE3	t-complex-associated-testis-expressed 3
219557	C7orf62	chromosome 7 open reading frame 62
8464	Supt3h	suppressor of Ty 3 homolog (S. cerevisiae)
6641	SNTB1	syntrophin, beta 1 (dystrophin-associated protein A1, 59kDa, basic component 1)
55780	C6orf70	chromosome 6 open reading frame 70
219578	ZNF804B	zinc finger protein 804B
102	ADAM10	ADAM metallopeptidase domain 10
253769	WDR27	WD repeat domain 27
79811	SLTM	SAFB-like, transcription modulator
166614	Dclk2	doublecortin-like kinase 2
55930	MYO5C	myosin VC
9355	Lhx2	LIM homeobox 2
54629	Fam63b	family with sequence similarity 63, member B
387263	C6orf120	chromosome 6 open reading frame 120
10681	GNB5	guanine nucleotide binding protein (G protein), beta 5
256130	TMEM196	transmembrane protein 196
257068	PHLDB2	pleckstrin homology-like domain, family B, member 2; phosphatidylinositol-specific phospholipase C, X domain containing 2
257068	PLCXD2	pleckstrin homology-like domain, family B, member 2; phosphatidylinositol-specific phospholipase C, X domain containing 2
90102	PHLDB2	pleckstrin homology-like domain, family B, member 2; phosphatidylinositol-specific phospholipase C, X domain containing 2
90102	PLCXD2	pleckstrin homology-like domain, family B, member 2; phosphatidylinositol-specific phospholipase C, X domain containing 2
136319	mtpn	myotrophin; leucine zipper protein 6
136319	LUZP6	myotrophin; leucine zipper protein 6
51187	RSL24D1	ribosomal L24 domain containing 1; similar to ribosomal protein L24-like
51187	RSL24D1P11	ribosomal L24 domain containing 1; similar to ribosomal protein L24-like
7342	ubp1	upstream binding protein 1 (LBP-1a)
5529	PPP2R5E	protein phosphatase 2, regulatory subunit B', epsilon isoform
54778	Rnf111	ring finger protein 111
23122	CLASP2	cytoplasmic linker associated protein 2
4644	MYO5A	myosin VA (heavy chain 12, myoxin)
8291	DYSF	dysferlin, limb girdle muscular dystrophy 2B (autosomal recessive)
Balkars		
Entrez ID	Official Gene Name	Gene Name

57630	sh3rf1	SH3 domain containing ring finger 1
2915	Grm5	glutamate receptor, metabotropic 5
5894	RAF1	v-raf-1 murine leukemia viral oncogene homolog 1
254827	Naaladl2	N-acetylated alpha-linked acidic dipeptidase-like 2
5915	Rarb	retinoic acid receptor, beta
55083	KIF26B	kinesin family member 26B
55287	TMEM40	transmembrane protein 40
134083	OR2Y1	olfactory receptor, family 2, subfamily Y, member 1
1089	CEACAM4	carcinoembryonic antigen-related cell adhesion molecule 4
83943	IMMP2L	IMP2 inner mitochondrial membrane peptidase-like (S. cerevisiae)
57585	Cramp1l	Crm, cramped-like (Drosophila)
8464	Supt3h	suppressor of Ty 3 homolog (S. cerevisiae)
56479	KCNQ5	potassium voltage-gated channel, KQT-like subfamily, member 5
6095	RORA	RAR-related orphan receptor A
124540	MSI2	musashi homolog 2 (Drosophila)
79652	TMEM204	transmembrane protein 204
1080	CFTR	cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)
9355	Lhx2	LIM homeobox 2
90861	HN1L	hematological and neurological expressed 1-like
285382	C3orf70	chromosome 3 open reading frame 70
60685	ZFAND3	zinc finger, AN1-type domain 3
2324	Flt4	fms-related tyrosine kinase 4
64478	CSMD1	CUB and Sushi multiple domains 1
25976	TIPARP	TCDD-inducible poly(ADP-ribose) polymerase
1087	CEACAM7	carcinoembryonic antigen-related cell adhesion molecule 7
23162	MAPK8IP3	mitogen-activated protein kinase 8 interacting protein 3
92483	LDHAL6B	lactate dehydrogenase A-like 6B
9742	IFT140	intraflagellar transport 140 homolog (Chlamydomonas)
92304	Scgb3a1	secretoglobin, family 3A, member 1
9215	LARGE	like-glycosyltransferase
7881	KCNAB1	potassium voltage-gated channel, shaker-related subfamily, beta member 1
4750	nek1	NIMA (never in mitosis gene a)-related kinase 1
8723	SNX4	sorting nexin 4
8411	EEA1	early endosome antigen 1
23609	mkrn2	makorin ring finger protein 2
4643	MYO1E	myosin IE
4684	Ncam1	neural cell adhesion molecule 1
90273	CEACAM21	carcinoembryonic antigen-related cell adhesion molecule 21
55768	ngly1	N-glycanase 1
57706	dennd1a	DENN/MADD domain containing 1A
9162	DGKI	diacylglycerol kinase, iota
6747	SSR3	signal sequence receptor, gamma (translocon-associated protein gamma)
8997	KALRN	kalirin, RhoGEF kinase
64764	CREB3L2	cAMP responsive element binding protein 3-like 2
57472	CNOT6	CCR4-NOT transcription complex, subunit 6
64084	Clstn2	calsyntenin 2
5629	PROX1	prospero homeobox 1
7155	TOP2B	topoisomerase (DNA) II beta 180kDa
1962	EHHADH	enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase
55101	ATP5SL	ATP5S-like
491	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2
5998	rgs3	regulator of G-protein signaling 3
9133	CCNB2	cyclin B2
7325	Ube2e2	ubiquitin-conjugating enzyme E2E 2 (UBC4/5 homolog, yeast)
114885	OSBPL11	oxysterol binding protein-like 11
136991	ASZ1	ankyrin repeat, SAM and basic leucine zipper domain containing 1
54674	lrrn3	leucine rich repeat neuronal 3
83992	Cttnbp2	cortactin binding protein 2
64283	RGNEF	Rho-guanine nucleotide exchange factor
Chechens		
Entrez ID	Official Gene Name	Gene Name
57538	ALPK3	alpha-kinase 3
220323	OAF	OAF homolog (Drosophila)
55846	ITFG2	integrin alpha FG-GAP repeat containing 2
157	Adrbk2	adrenergic, beta, receptor kinase 2
23650	TRIM29	tripartite motif-containing 29
2288	fkbp4	FK506 binding protein 4, 59kDa
23558	WBP2	WW domain binding protein 2

57728	WDR19	WD repeat domain 19
84002	b3gnt5	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5
4756	NEO1	neogenin homolog 1 (chicken)
1089	CEACAM4	carcinoembryonic antigen-related cell adhesion molecule 4
9696	Crocc	ciliary rootlet coiled-coil, rootletin
51088	klhl5	kelch-like 5 (Drosophila)
6390	SdhB	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
79783	C7orf10	chromosome 7 open reading frame 10
4828	NMB	neuromedin B
23400	Atp13a2	ATPase type 13A2
4488	MSX2	msh homeobox 2
83714	NRIP2	nuclear receptor interacting protein 2
5602	MAPK10	mitogen-activated protein kinase 10
831	cast	calpastatin
23704	KCNE4	potassium voltage-gated channel, Isk-related family, member 4
9252	RPS6KA5	ribosomal protein S6 kinase, 90kDa, polypeptide 5
55917	CTTNBP2NL	CTTNBP2 N-terminal like
60685	ZFAND3	zinc finger, AN1-type domain 3
775	LOC100131098	hypothetical protein LOC100131098; calcium channel, voltage-dependent, L type, alpha 1C subunit
775	CACNA1C	hypothetical protein LOC100131098; calcium channel, voltage-dependent, L type, alpha 1C subunit
257068	PHLDB2	pleckstrin homology-like domain, family B, member 2; phosphatidylinositol-specific phospholipase C, X domain containing 2
257068	PLCXD2	pleckstrin homology-like domain, family B, member 2; phosphatidylinositol-specific phospholipase C, X domain containing 2
7289	TULP3	tubby like protein 3
64478	CSMD1	CUB and Sushi multiple domains 1
84448	ABLIM2	actin binding LIM protein family, member 2
1087	CEACAM7	carcinoembryonic antigen-related cell adhesion molecule 7
137868	SGCZ	sarcoglycan zeta
57520	Hecw2	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2
91107	TRIM47	tripartite motif-containing 47
90102	PHLDB2	pleckstrin homology-like domain, family B, member 2; phosphatidylinositol-specific phospholipase C, X domain containing 2
90102	PLCXD2	pleckstrin homology-like domain, family B, member 2; phosphatidylinositol-specific phospholipase C, X domain containing 2
144347	fam101a	family with sequence similarity 101, member A
23478	SEC11A	SEC11 homolog A (S. cerevisiae)
64978	Mrpl38	mitochondrial ribosomal protein L38
8723	SNX4	sorting nexin 4
201292	Trim65	tripartite motif-containing 65
11240	PADI2	peptidyl arginine deiminase, type II
25833	POU2F3	POU class 2 homeobox 3
201294	UNC13D	unc-13 homolog D (C. elegans)
9640	ZNF592	zinc finger protein 592
3211	HOXB1	homeobox B1
3218	HOXB8	homeobox B8
3219	HOXB9	homeobox B9
3216	HOXB6	homeobox B6
91355	LRP5L	low density lipoprotein receptor-related protein 5-like
3217	HOXB7	homeobox B7
3214	HOXB4	homeobox B4
60495	Hpse2	heparanase 2
90273	CEACAM21	carcinoembryonic antigen-related cell adhesion molecule 21
3215	hoxb5	homeobox B5
3212	HOXB2	homeobox B2
166647	GPR125	G protein-coupled receptor 125
3213	HOXB3	homeobox B3
9262	STK17B	serine/threonine kinase 17b
51752	ERAP1	endoplasmic reticulum aminopeptidase 1
100134934	C17orf106	hypothetical protein LOC100134934
219902	Tmem136	transmembrane protein 136
3751	KCND2	potassium voltage-gated channel, Shal-related subfamily, member 2
85302	fbf1	Fas (TNFRSF6) binding factor 1
84366	PRAC	prostate cancer susceptibility candidate
3756	KCNH1	potassium voltage-gated channel, subfamily H (eag-related), member 1
80008	Tmem156	transmembrane protein 156
4237	MFAP2	microfibrillar-associated protein 2

64084	Clstn2	calsyntenin 2
55101	ATP5SL	ATP5S-like
2305	FOXM1	forkhead box M1
1018	Cdk3	cyclin-dependent kinase 3
55733	Hhat	hedgehog acyltransferase
60312	AFAP1	actin filament associated protein 1
114885	OSBPL11	oxysterol binding protein-like 11
51	ACOX1	acyl-Coenzyme A oxidase 1, palmitoyl
7225	Trpc6	transient receptor potential cation channel, subfamily C, member 6
85451	unk	unkempt homolog (Drosophila)
23101	MCF2L2	MCF.2 cell line derived transforming sequence-like 2
2181	ACSL3	acyl-CoA synthetase long-chain family member 3
80017	C14orf159	chromosome 14 open reading frame 159
Georgians		
Entrez ID	Official Gene Name	Gene Name
57630	sh3rf1	SH3 domain containing ring finger 1
84930	MASTL	microtubule associated serine/threonine kinase-like
1761	dmrt1	doublesex and mab-3 related transcription factor 1
84529	c15orf41	chromosome 15 open reading frame 41
112849	C14orf149	chromosome 14 open reading frame 149
5578	Prkca	protein kinase C, alpha
23233	EXOC6B	exocyst complex component 6B
149699	GTSF1L	gametocyte specific factor 1-like
56829	ZC3HAV1	zinc finger CCCH-type, antiviral 1
157	Adrbk2	adrenergic, beta, receptor kinase 2
64582	GPR135	G protein-coupled receptor 135
54906	C10orf18	chromosome 10 open reading frame 18
4756	NEO1	neogenin homolog 1 (chicken)
8833	GMPS	guanine monphosphate synthetase
9696	Crocc	ciliary rootlet coiled-coil, rootletin
6259	RYK	RYK receptor-like tyrosine kinase
57670	KIAA1549	KIAA1549
6390	SdhB	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
9595	CYTIP	cytohesin 1 interacting protein
219557	C7orf62	chromosome 7 open reading frame 62
9650	MTFR1	mitochondrial fission regulator 1
144577	c12orf66	chromosome 12 open reading frame 66
23400	Atp13a2	ATPase type 13A2
51528	JKAMP	chromosome 14 open reading frame 100
124540	MSI2	musashi homolog 2 (Drosophila)
79754	ASB13	ankyrin repeat and SOCS box-containing 13
64478	CSMD1	CUB and Sushi multiple domains 1
92092	ZC3HAV1L	zinc finger CCCH-type, antiviral 1-like
56164	STK31	serine/threonine kinase 31
57522	SRGAP1	SLIT-ROBO Rho GTPase activating protein 1
1794	DOCK2	dedicator of cytokinesis 2
51098	IFT52	intraflagellar transport 52 homolog (Chlamydomonas)
3918	LAMC2	laminin, gamma 2
4605	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2
8788	DLK1	delta-like 1 homolog (Drosophila)
7881	KCNAB1	potassium voltage-gated channel, shaker-related subfamily, beta member 1
57596	BEGAIN	brain-enriched guanylate kinase-associated homolog (rat)
11240	PADI2	peptidyl arginine deiminase, type II
100131897	FAM196B	Uncharacterized protein LOC100131897
11081	Kera	keratocan
55156	ARMC1	armadillo repeat containing 1
91355	LRP5L	low density lipoprotein receptor-related protein 5-like
114801	TMEM200A	transmembrane protein 200A
27145	FILIP1	filamin A interacting protein 1
79977	GRHL2	grainyhead-like 2 (Drosophila)
10730	YME1L1	YME1-like 1 (S. cerevisiae)
3751	KCND2	potassium voltage-gated channel, Shal-related subfamily, member 2
83988	NCALD	neurocalcin delta
23294	Anks1a	ankyrin repeat and sterile alpha motif domain containing 1A
9197	slc33a1	solute carrier family 33 (acetyl-CoA transporter), member 1
222663	SCUBE3	signal peptide, CUB domain, EGF-like 3
10110	SGK2	serum/glucocorticoid regulated kinase 2
219578	ZNF804B	zinc finger protein 804B
23057	Nmnat2	nicotinamide nucleotide adenylyltransferase 2

91452	acbd5	acyl-Coenzyme A binding domain containing 5
4091	SMAD6	SMAD family member 6
9890	LPPR4	plasticity related gene 1
4237	MFAP2	microfibrillar-associated protein 2
27091	CACNG5	calcium channel, voltage-dependent, gamma subunit 5
27092	Cacng4	calcium channel, voltage-dependent, gamma subunit 4
729665	C14orf38	chromosome 14 open reading frame 38
5629	PROX1	prospero homeobox 1
130399	ACVR1C	activin A receptor, type IC
285315	c3orf33	chromosome 3 open reading frame 33
23002	DAAM1	dishevelled associated activator of morphogenesis 1
9760	Tox	thymocyte selection-associated high mobility group box
5998	rgs3	regulator of G-protein signaling 3
6954	Tcp11	t-complex 11 homolog (mouse)
154075	SAMD3	sterile alpha motif domain containing 3
4060	LUM	lumican
58524	Dmrt3	doublesex and mab-3 related transcription factor 3
23007	PLCH1	phospholipase C, eta 1
1634	DCN	decorin
8848	Tsc22d1	TSC22 domain family, member 1
Lezgins_NorthOssetians		
Entrez ID	Official Gene Name	Gene Name
57630	sh3rf1	SH3 domain containing ring finger 1
7498	XDH	xanthine dehydrogenase
5578	Prkca	protein kinase C, alpha
254827	Naaladl2	N-acetylated alpha-linked acidic dipeptidase-like 2
5915	Rarb	retinoic acid receptor, beta
285	ANGPT2	angiopoietin 2
79648	MCPH1	microcephalin 1
1089	CEACAM4	carcinoembryonic antigen-related cell adhesion molecule 4
9046	DOK2	docking protein 2, 56kDa
1244	Abcc2	ATP-binding cassette, sub-family C (CFTR/MRP), member 2
2675	LOC100132336	similar to GDNF family receptor alpha 2; GDNF family receptor alpha 2
2675	GFRA2	similar to GDNF family receptor alpha 2; GDNF family receptor alpha 2
63892	Thada	thyroid adenoma associated
9595	CYTIP	cytohesin 1 interacting protein
23039	XPO7	exportin 7
10659	CELF2	CUG triplet repeat, RNA binding protein 2
64478	CSMD1	CUB and Sushi multiple domains 1
1087	CEACAM7	carcinoembryonic antigen-related cell adhesion molecule 7
23268	DNMBP	dynamin binding protein
55806	HR	hairless homolog (mouse)
54926	ube2r2	ubiquitin-conjugating enzyme E2R 2
8822	FGF17	fibroblast growth factor 17
4750	nek1	NIMA (never in mitosis gene a)-related kinase 1
55691	Frmd4a	FERM domain containing 4A
6444	SGCD	sarcoglycan, delta (35kDa dystrophin-associated glycoprotein)
54439	RBM27	RNA binding motif protein 27
7103	Tspan8	tetraspanin 8
90273	CEACAM21	carcinoembryonic antigen-related cell adhesion molecule 21
60495	Hpse2	heparanase 2
22998	LIMCH1	LIM and calponin homology domains 1
55768	ngly1	N-glycanase 1
27145	FILIP1	filamin A interacting protein 1
5711	PSMD5	proteasome (prosome, macropain) 26S subunit, non-ATPase, 5
5459	POU4F3	POU class 4 homeobox 3
55833	UBAP2	ubiquitin associated protein 2
3751	KCND2	potassium voltage-gated channel, Shal-related subfamily, member 2
727	C5	complement component 5
10361	Npm2	nucleophosmin/nucleoplasmin, 2
130271	PLEKHH2	pleckstrin homology domain containing, family H (with MyTH4 domain) member 2
26147	phf19	PHD finger protein 19
7185	Traf1	TNF receptor-associated factor 1
3257	Hps1	Hermansky-Pudlak syndrome 1
27091	CACNG5	calcium channel, voltage-dependent, gamma subunit 5
64760	FAM160B2	family with sequence similarity 160, member B2
27092	Cacng4	calcium channel, voltage-dependent, gamma subunit 4
7155	TOP2B	topoisomerase (DNA) II beta 180kDa
79873	Nudt18	nudix (nucleoside diphosphate linked moiety X)-type motif 18

55101	ATP5SL	ATP5S-like
130399	ACVR1C	activin A receptor, type IC
80346	REEP4	receptor accessory protein 4
5529	PPP2R5E	protein phosphatase 2, regulatory subunit B', epsilon isoform
6862	t	T, brachyury homolog (mouse)
2039	epb49	erythrocyte membrane protein band 4.9 (dematin)

7.2 Candidate gene list for high altitude adaptation

Supplementary Table 2 Candidate gene list for high altitude adaptation screening.

GO term "oxygen homeostasis"		
Entrez ID	Official gene name	Gene name
7477	WNT7B	wingless-type MMTV integration site family, member 7B
212	ALAS2	aminolevulinate, delta-, synthase 2
3091	HIF1A	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
64428	NARFL	nuclear prelamin A recognition factor-like
54583	EGLN1	egl nine homolog 1 (C. elegans)
6648	SOD2	superoxide dismutase 2, mitochondrial
List adopted from Pagani et al. (2012)		
Entrez ID	Official gene name	Gene name
5465	PPARA	peroxisome proliferator-activated receptor alpha
112399	egln3	egl nine homolog 3 (C. elegans)
2034	EPAS1	endothelial PAS domain protein 1
54583	egln1	egl nine homolog 1 (C. elegans)
1906	EDN1	endothelin 1
112398	EGLN2	egl nine homolog 2 (C. elegans)
1636	ACE	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1
7422	VEGFA	vascular endothelial growth factor A
7428	vhl	von Hippel-Lindau tumor suppressor
4846	NOS3	nitric oxide synthase 3 (endothelial cell)
2057	EPOR	erythropoietin receptor
3047	HBG1	hemoglobin, gamma A
5230	PGK1	phosphoglycerate kinase 1
3045	hbd	hemoglobin, delta
3043	HBB	hemoglobin, beta
2056	EPO	erythropoietin
GO term "hypoxia response"		
Entrez ID	Official gene name	gene name
847	cat	catalase
384	arg2	arginase, type II
9915	ARNT2	aryl-hydrocarbon receptor nuclear translocator 2
100133941	CD24L4	CD24 molecule; CD24 molecule-like 4
938	CD24L4	CD24 molecule; CD24 molecule-like 4
230	aldoc	aldolase C, fructose-bisphosphate
57085	Agtrap	angiotensin II receptor-associated protein
664	BNIP3	BCL2/adenovirus E1B 19kDa interacting protein 3
481	Atp1b1	ATPase, Na+/K+ transporting, beta 1 polypeptide
1103	chat	choline acetyltransferase
405	ARNT	aryl hydrocarbon receptor nuclear translocator
218	ALDH3A1	aldehyde dehydrogenase 3 family, memberA1
1636	ACE	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1
6347	Ccl2	chemokine (C-C motif) ligand 2
857	cav1	caveolin 1, caveolae protein, 22kDa
650	bmp2	bone morphogenetic protein 2
1137	Chrna4	cholinergic receptor, nicotinic, alpha 4
596	BCL2	B-cell CLL/lymphoma 2
134	Adora1	adenosine A1 receptor
285	ANGPT2	angiopoietin 2
952	cd38	CD38 molecule
6868	Adam17	ADAM metallopeptidase domain 17
51129	ANGPTL4	angiopoietin-like 4
81	actn4	actinin, alpha 4
651610	LOC651610	similar to Serine-protein kinase ATM (Ataxia telangiectasia mutated) (A-T, mutated); ataxia telangiectasia mutated
472	LOC651610	similar to Serine-protein kinase ATM (Ataxia telangiectasia mutated) (A-T, mutated); ataxia telangiectasia mutated
81575	APOLD1	apolipoprotein L domain containing 1
100133941	Cd24	CD24 molecule; CD24 molecule-like 4
938	Cd24	CD24 molecule; CD24 molecule-like 4
824	CAPN2	calpain 2, (m/II) large subunit
9370	ADIPOQ	adiponectin, C1Q and collagen domain containing
651610	ATM	similar to Serine-protein kinase ATM (Ataxia telangiectasia mutated) (A-T,

Supplementary Table 2

		mutated); ataxia telangiectasia mutated
472	ATM	similar to Serine-protein kinase ATM (Ataxia telangiectasia mutated) (A-T, mutated); ataxia telangiectasia mutated
212	alaS2	aminolevulinate, delta-, synthase 2
100	ada	adenosine deaminase
430	ASCL2	achaete-scute complex homolog 2 (Drosophila)
283	Ang	angiogenin, ribonuclease, RNase A family, 5
133	ADM	adrenomedullin
18	ABAT	4-aminobutyrate aminotransferase
158	ADSL	adenylosuccinate lyase
GO term "oxygen transport"		
Entrez ID	Official gene name	Gene name
3050	HBZ	hemoglobin, zeta
79747	C6orf103	chromosome 6 open reading frame 103
3049	HBQ1	hemoglobin, theta 1
26034	IPCEF1	interaction protein for cytohesin exchange factors 1
114757	CYGB	cytoglobin
58157	NGB	neuroglobin
3039	HBA1	hemoglobin, alpha 2; hemoglobin, alpha 1
3040	HBA1	hemoglobin, alpha 2; hemoglobin, alpha 1
3046	HBE1	hemoglobin, epsilon 1
3043	HBB	hemoglobin, beta
4609	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
selected genes in Tibetans		
Entrez ID	Official gene name	gene name
5465	PPARA	peroxisome proliferator-activated receptor alpha
2034	EPAS1	endothelial PAS domain protein 1
54583	EGLN1	egl nine homolog 1 (C. elegans)
1571	CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1
5728	PTEN	phosphatase and tensin homolog; phosphatase and tensin homolog pseudogene 1
11191	PTEN	phosphatase and tensin homolog; phosphatase and tensin homolog pseudogene 1
1909	EDNRA	endothelin receptor type A
1586	CYP17A1	cytochrome P450, family 17, subfamily A, polypeptide 1
3162	HMOX1	heme oxygenase (decycling) 1
817	CAMK2D	calcium/calmodulin-dependent protein kinase II delta
51129	ANGPTL4	angiopoietin-like 4
selected genes in Ethiopians		
Entrez ID	Official gene name	gene name
10367	cbara1	calcium binding atopy-related autoantigen 1
10451	vav3	vav 3 guanine nucleotide exchange factor
7068	thrb	thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian)

7.3 Candidate gene list for longevity screening

Supplementary Table 3 Candidate gene list for longevity screening

GO term "aging"		
Entrez ID	Official gene name	Gene name
6722	Srf	serum response factor (c-fos serum response element-binding transcription factor)
646359	LOC646316	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene
23411	sirt1	sirtuin (silent mating type information regulation 2 homolog) 1 (S. cerevisiae)
27250	PDCD4	programmed cell death 4 (neoplastic transformation inhibitor)
3265	hras	v-Ha-ras Harvey rat sarcoma viral oncogene homolog
6926	tbx3	T-box 3
6909	tbx2	T-box 2
675	Brca2	breast cancer 2, early onset
729686	LOC399804	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 21; hypothetical LOC100131044; similar to nucleophosmin 1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)
1029	CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
646359	LOC646127	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene
729686	Npm1	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 21; hypothetical LOC100131044; similar to nucleophosmin 1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)
3611	ILK	integrin-linked kinase
286053	NSMCE2	non-SMC element 2, MMS21 homolog (S. cerevisiae)
50507	Nox4	NADPH oxidase 4
3383	ICAM1	intercellular adhesion molecule 1
23636	Nup62	nucleoporin 62kDa
729686	LOC729686	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 21; hypothetical LOC100131044; similar to nucleophosmin 1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)
5580	PRKCD	protein kinase C, delta
7486	wrn	similar to Werner syndrome protein; Werner syndrome, RecQ helicase-like
652522	wrn	similar to Werner syndrome protein; Werner syndrome, RecQ helicase-like
729686	LOC100131044	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 21; hypothetical LOC100131044; similar to nucleophosmin 1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)
3028	HSD17B10	hydroxysteroid (17-beta) dehydrogenase 10
3987	LIMS1	LIM and senescent cell antigen-like domains 1
8626	tp63	tumor protein p63
729686	LOC729342	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 21; hypothetical LOC100131044; similar to nucleophosmin 1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)
652346	PML	promyelocytic leukemia; similar to promyelocytic leukemia protein isoform 1
5371	PML	promyelocytic leukemia; similar to promyelocytic leukemia protein isoform 1
4335	mnt	MAX binding protein
811	CALR	calreticulin
23515	MORC3	MORC family CW-type zinc finger 3
652346	LOC652346	promyelocytic leukemia; similar to promyelocytic leukemia protein isoform 1
5371	LOC652346	promyelocytic leukemia; similar to promyelocytic leukemia protein isoform 1
596	BCL2	B-cell CLL/lymphoma 2
11243	pmf1	bone gamma-carboxyglutamate (gla) protein; polyamine-modulated factor 1
632	pmf1	bone gamma-carboxyglutamate (gla) protein; polyamine-modulated factor 1
729686	LOC440577	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 21; hypothetical LOC100131044; similar to nucleophosmin 1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)
9093	DNAJA3	DnaJ (Hsp40) homolog, subfamily A, member 3
7014	TERF2	telomeric repeat binding factor 2
646359	TERF1	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene
983	Cdk1	cell division cycle 2, G1 to S and G2 to M
7157	tp53	tumor protein p53
11243	BGLAP	bone gamma-carboxyglutamate (gla) protein; polyamine-modulated factor 1

632	BGLAP	bone gamma-carboxyglutamate (gla) protein; polyamine-modulated factor 1
5604	MAP2K1	mitogen-activated protein kinase kinase 1
3064	HTT	huntingtin
729686	NPM1P21	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 21; hypothetical LOC100131044; similar to nucleophosmin 1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)
79677	SMC6	structural maintenance of chromosomes 6
7486	LOC652522	similar to Werner syndrome protein; Werner syndrome, RecQ helicase-like
652522	LOC652522	similar to Werner syndrome protein; Werner syndrome, RecQ helicase-like
6647	SOD1	superoxide dismutase 1, soluble
1026	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
3398	ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
23137	smc5	structural maintenance of chromosomes 5
84417	c2orf40	chromosome 2 open reading frame 40
646359	LOC283523	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene
GO term "cell senescence"		
Entrez ID	Official gene name	gene name
10783	Nek6	NIMA (never in mitosis gene a)-related kinase 6
11200	LOC100133012	protein kinase CHK2-like; CHK2 checkpoint homolog (S. pombe); similar to hCG1983233
646096	LOC100133012	protein kinase CHK2-like; CHK2 checkpoint homolog (S. pombe); similar to hCG1983233
100133012	LOC100133012	protein kinase CHK2-like; CHK2 checkpoint homolog (S. pombe); similar to hCG1983233
3622	ING2	inhibitor of growth family, member 2
6722	Srf	serum response factor (c-fos serum response element-binding transcription factor)
9891	NUAK1	NUAK family, SNF1-like kinase, 1
8550	MAPKAPK5	mitogen-activated protein kinase-activated protein kinase 5
23411	sirt1	sirtuin (silent mating type information regulation 2 homolog) 1 (S. cerevisiae)
4311	MME	membrane metallo-endopeptidase
1111	CHEK1	CHK1 checkpoint homolog (S. pombe)
11200	CHEK2	protein kinase CHK2-like; CHK2 checkpoint homolog (S. pombe); similar to hCG1983233
811	CALR	calreticulin
1432	Mapk14	mitogen-activated protein kinase 14
3265	hras	v-Ha-ras Harvey rat sarcoma viral oncogene homolog
87178	Pnpt1	polyribonucleotide nucleotidyltransferase 1
651921	LOC648152	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
648152	LOC648152	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
545	LOC648152	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
1029	CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
286053	NSMCE2	non-SMC element 2, MMS21 homolog (S. cerevisiae)
22925	PLA2R1	phospholipase A2 receptor 1, 180kDa
25989	ulk3	unc-51-like kinase 3 (C. elegans)
7014	TERF2	telomeric repeat binding factor 2
7015	TERT	telomerase reverse transcriptase
7291	twist1	twist homolog 1 (Drosophila)
651610	LOC651610	similar to Serine-protein kinase ATM (Ataxia telangiectasia mutated) (A-T, mutated); ataxia telangiectasia mutated
472	LOC651610	similar to Serine-protein kinase ATM (Ataxia telangiectasia mutated) (A-T, mutated); ataxia telangiectasia mutated
100130009	LOC100130009	hypothetical LOC100130009; high mobility group AT-hook 1
3159	LOC100130009	hypothetical LOC100130009; high mobility group AT-hook 1
7157	tp53	tumor protein p53
5604	MAP2K1	mitogen-activated protein kinase kinase 1
79677	SMC6	structural maintenance of chromosomes 6
1021	Cdk6	cyclin-dependent kinase 6
651921	LOC651921	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
648152	LOC651921	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
545	LOC651921	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
51384	Wnt16	wingless-type MMTV integration site family, member 16

Supplementary Table 3

651921	ATR	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
648152	ATR	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
545	ATR	ataxia telangiectasia and Rad3 related; similar to ataxia telangiectasia and Rad3 related protein
8091	HMGA2	high mobility group AT-hook 2
5580	PRKCD	protein kinase C, delta
100130009	HMGA1	hypothetical LOC100130009; high mobility group AT-hook 1
3159	HMGA1	hypothetical LOC100130009; high mobility group AT-hook 1
651610	ATM	similar to Serine-protein kinase ATM (Ataxia telangiectasia mutated) (A-T, mutated); ataxia telangiectasia mutated
472	ATM	similar to Serine-protein kinase ATM (Ataxia telangiectasia mutated) (A-T, mutated); ataxia telangiectasia mutated
1026	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
3398	ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
11200	LOC646096	protein kinase CHK2-like; CHK2 checkpoint homolog (S. pombe); similar to hCG1983233
646096	LOC646096	protein kinase CHK2-like; CHK2 checkpoint homolog (S. pombe); similar to hCG1983233
100133012	LOC646096	protein kinase CHK2-like; CHK2 checkpoint homolog (S. pombe); similar to hCG1983233
23137	smc5	structural maintenance of chromosomes 5
84417	c2orf40	chromosome 2 open reading frame 40
GO term "telomer maintenance"		
Entrez ID	Official gene name	Gene name
10728	PTGES3	prostaglandin E synthase 3 (cytosolic)
4683	NBN	nibrin
5422	POLA1	polymerase (DNA directed), alpha 1, catalytic subunit
731751	PRKDC	similar to protein kinase, DNA-activated, catalytic polypeptide; protein kinase, DNA-activated, catalytic polypeptide
5591	PRKDC	similar to protein kinase, DNA-activated, catalytic polypeptide; protein kinase, DNA-activated, catalytic polypeptide
54386	TERF2IP	telomeric repeat binding factor 2, interacting protein
23649	POLA2	polymerase (DNA directed), alpha 2 (70kD subunit)
25913	POT1	POT1 protection of telomeres 1 homolog (S. pombe)
6119	RPA3	replication protein A3, 14kDa
5557	PRIM1	primase, DNA, polypeptide 1 (49kDa)
6117	RPA1	replication protein A1, 70kDa
6118	RPA2	replication protein A2, 32kDa
65057	ACD	adrenocortical dysplasia homolog (mouse)
1736	DKC1	dyskeratosis congenita 1, dyskerin
5427	POLE2	polymerase (DNA directed), epsilon 2 (p59 subunit)
5558	PRIM2	primase, DNA, polypeptide 2 (58kDa)
26272	FBXO4	F-box protein 4
8290	HIST3H3	histone cluster 3, H3
2072	ERCC4	excision repair cross-complementing rodent repair deficiency, complementation group 4
26277	TINF2	TERF1 (TRF1)-interacting nuclear factor 2
8370	HIST1H4J	histone cluster 1, H4l; histone cluster 1, H4k; histone cluster 4, H4; histone cluster 1, H4h; histone cluster 1, H4j; histone cluster 1, H4i; histone cluster 1, H4d; histone cluster 1, H4c; histone cluster 1, H4f; histone cluster 1, H4e; histone cluster 1, H4b; histone cluster 1, H4a; histone cluster 2, H4a; histone cluster 2, H4b
7014	TERF2	telomeric repeat binding factor 2
7015	TERT	telomerase reverse transcriptase
8771	RTEL1	tumor necrosis factor receptor superfamily, member 6b, decoy; regulator of telomere elongation helicase 1
51750	RTEL1	tumor necrosis factor receptor superfamily, member 6b, decoy; regulator of telomere elongation helicase 1
2237	FEN1	flap structure-specific endonuclease 1
646359	TERF1	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene
646127	TERF1	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene
283523	TERF1	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene

Supplementary Table 3

7013	TERF1	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene
646316	TERF1	similar to telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1; telomeric repeat binding factor (NIMA-interacting) 1 pseudogene
23293	SMG6	Smg-6 homolog, nonsense mediated mRNA decay factor (C. elegans)
3978	LIG1	ligase I, DNA, ATP-dependent
5426	POLE	polymerase (DNA directed), epsilon
10111	RAD50	RAD50 homolog (S. cerevisiae)
10714	POLD3	polymerase (DNA-directed), delta 3, accessory subunit
5985	RFC5	replication factor C (activator 1) 5, 36.5kDa
64421	DCLRE1C	DNA cross-link repair 1C (PSO2 homolog, S. cerevisiae)
1763	DNA2	DNA replication helicase 2 homolog (yeast)
57804	POLD4	polymerase (DNA-directed), delta 4
5983	RFC3	replication factor C (activator 1) 3, 38kDa
64858	DCLRE1B	DNA cross-link repair 1B (PSO2 homolog, S. cerevisiae)
5984	RFC4	replication factor C (activator 1) 4, 37kDa
5981	RFC1	replication factor C (activator 1) 1, 145kDa
5982	RFC2	replication factor C (activator 1) 2, 40kDa
5424	POLD1	polymerase (DNA directed), delta 1, catalytic subunit 125kDa
79991	OBFC1	oligonucleotide/oligosaccharide-binding fold containing 1
5425	POLD2	polymerase (DNA directed), delta 2, regulatory subunit 50kDa
5111	PCNA	proliferating cell nuclear antigen
142	PARP1	poly (ADP-ribose) polymerase 1

7.4 List of DAVID gene enrichments

Supplementary Table 4 List of DAVID gene enrichments of Abhkazians and Kumyks_Adygei. Enrichment were carried out with the genes form Supplementary Table 1. Other enrichments are not listed because of lack of space.

Abhkazians			
Annotation Cluster 1	Enrichment Score: 1.2882573285581826		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0060538~skeletal muscle organ development	0,022	CAV2, CAV1, RXRG
GOTERM_BP_FAT	GO:0007519~skeletal muscle tissue development	0,022	CAV2, CAV1, RXRG
GOTERM_BP_FAT	GO:0014706~striated muscle tissue development	0,064	CAV2, CAV1, RXRG
GOTERM_BP_FAT	GO:0060537~muscle tissue development	0,070	CAV2, CAV1, RXRG
GOTERM_BP_FAT	GO:0007517~muscle organ development	0,166	CAV2, CAV1, RXRG
Annotation Cluster 2	Enrichment Score: 1.0853554717838654		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0007017~microtubule-based process	0,002	CAV2, CAV1, HAUS5, TUBB2B, UBE2C, KIFC3
GOTERM_BP_FAT	GO:0000226~microtubule cytoskeleton organization	0,014	CAV2, CAV1, HAUS5, UBE2C
GOTERM_BP_FAT	GO:0007010~cytoskeleton organization	0,017	CAV2, CAV1, HAUS5, RAF1, NEURL2, UBE2C
KEGG_PATHWAY	hsa04510:Focal adhesion	0,172	CAV2, CAV1, RAF1
GOTERM_BP_FAT	GO:0043086~negative regulation of catalytic activity	0,250	CAV1, RGS3, UBE2C
GOTERM_BP_FAT	GO:0044092~negative regulation of molecular function	0,324	CAV1, RGS3, UBE2C
GOTERM_BP_FAT	GO:0007242~intracellular signaling cascade	0,445	CAV1, RGS3, CHN2, RAF1, NEURL2, UBE2C
GOTERM_CC_FAT	GO:0005829~cytosol	0,765	CAV2, CAV1, RGS3, RAF1, UBE2C
Annotation Cluster 3	Enrichment Score: 1.0737151260922138		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0007010~cytoskeleton organization	0,017	CAV2, CAV1, HAUS5, RAF1, NEURL2, UBE2C
GOTERM_CC_FAT	GO:0005924~cell-substrate adherens junction	0,062	CAV2, CAV1, NEURL2
GOTERM_CC_FAT	GO:0030055~cell-substrate junction	0,069	CAV2, CAV1, NEURL2
GOTERM_CC_FAT	GO:0005912~adherens junction	0,119	CAV2, CAV1, NEURL2
GOTERM_CC_FAT	GO:0030054~cell junction	0,136	CAV2, CAV1, NLGN1, LIN7C, NEURL2
GOTERM_CC_FAT	GO:0070161~anchoring junction	0,141	CAV2, CAV1, NEURL2
GOTERM_CC_FAT	GO:0016323~basolateral plasma membrane	0,183	CAV2, CAV1, NEURL2
Annotation Cluster 4	Enrichment Score: 1.037362447208418		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0030317~sperm motility	0,003	GAPDHS, CAV2, SLC9A10
GOTERM_CC_FAT	GO:0019861~flagellum	0,013	GAPDHS, CAV2, SLC9A10
GOTERM_BP_FAT	GO:0032504~multicellular organism reproduction	0,088	GAPDHS, CAV2, ACOX1, CAV1, SLC9A10
GOTERM_BP_FAT	GO:0048609~reproductive process in a multicellular organism	0,088	GAPDHS, CAV2, ACOX1, CAV1, SLC9A10
GOTERM_BP_FAT	GO:0006928~cell motion	0,228	GAPDHS, CAV2, NRXN1, SLC9A10
GOTERM_BP_FAT	GO:0051674~localization of cell	0,289	GAPDHS, CAV2, SLC9A10
GOTERM_BP_FAT	GO:0048870~cell motility	0,289	GAPDHS, CAV2, SLC9A10
GOTERM_CC_FAT	GO:0042995~cell projection	0,755	GAPDHS, CAV2, SLC9A10
Annotation Cluster 5	Enrichment Score: 0.969298398825397		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0016192~vesicle-mediated transport	0,049	CAV2, CAV1, UNC13D, NLGN1, LIN7C, DOPEY1
GOTERM_BP_FAT	GO:0006887~exocytosis	0,060	UNC13D, NLGN1, LIN7C
GOTERM_BP_FAT	GO:0032940~secretion by cell	0,162	UNC13D, NLGN1, LIN7C
GOTERM_BP_FAT	GO:0046903~secretion	0,280	UNC13D, NLGN1, LIN7C
Annotation Cluster 6	Enrichment Score: 0.8700319160801923		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0007017~microtubule-based process	0,002	CAV2, CAV1, HAUS5, TUBB2B, UBE2C, KIFC3
GOTERM_BP_FAT	GO:0051258~protein polymerization	0,013	CAV2, CAV1, TUBB2B
GOTERM_BP_FAT	GO:0043623~cellular protein complex assembly	0,109	CAV2, CAV1, TUBB2B
GOTERM_BP_FAT	GO:0051259~protein oligomerization	0,122	CAV2, CAV1, UPK1A
GOTERM_BP_FAT	GO:0006461~protein complex assembly	0,256	CAV2, CAV1, TUBB2B, UPK1A
GOTERM_BP_FAT	GO:0070271~protein complex biogenesis	0,256	CAV2, CAV1, TUBB2B, UPK1A
GOTERM_BP_FAT	GO:0034622~cellular macromolecular complex assembly	0,303	CAV2, CAV1, TUBB2B
GOTERM_BP_FAT	GO:0034621~cellular macromolecular complex subunit organization	0,353	CAV2, CAV1, TUBB2B
GOTERM_BP_FAT	GO:0065003~macromolecular complex assembly	0,409	CAV2, CAV1, TUBB2B, UPK1A

GOTERM_BP_FAT	GO:0043933~macromolecular complex subunit organization	0,451	CAV2, CAV1, TUBB2B, UPK1A
GOTERM_MF_FAT	GO:0042802~identical protein binding	0,703	CAV2, CAV1, UPK1A
Annotation Cluster 7	Enrichment Score: 0.83543181824248		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0005739~mitochondrion	0,021	CAV2, ACOX1, CAV1, ACOT8, GCET2, COX6B1, RAF1, CTSA, MRPL38, MACROD1
GOTERM_CC_FAT	GO:0031090~organelle membrane	0,241	CAV2, ACOX1, CAV1, LARGE, COX6B1, RAF1, DOPEY1
SP_PIR_KEYWORDS	disease mutation	0,621	ACOX1, CAV1, LARGE, COX6B1, RAF1, CTSA, CNGB1
Annotation Cluster 8	Enrichment Score: 0.72728993306235		
Category	Term	PValue	Genes
INTERPRO	IPR017907:Zinc finger, RING-type, conserved site	0,100	TRIM65, TRIM47, UNK, MKRN2
INTERPRO	IPR001841:Zinc finger, RING-type	0,113	TRIM65, TRIM47, UNK, MKRN2
SMART	SM00184:RING	0,141	TRIM65, TRIM47, UNK, MKRN2
SP_PIR_KEYWORDS	zinc-finger	0,147	TRIM65, TRIM47, ZSWIM1, ZNF385D, RXRG, CHN2, RAF1, UNK, ZSWIM3, ZNF335, MKRN2
UP_SEQ_FEATURE	zinc finger region:RING-type	0,203	TRIM65, TRIM47, MKRN2
INTERPRO	IPR018957:Zinc finger, C3HC4 RING-type	0,228	TRIM65, TRIM47, MKRN2
GOTERM_MF_FAT	GO:0008270~zinc ion binding	0,372	TRIM65, TRIM47, ZSWIM1, ZNF385D, RXRG, CHN2, RAF1, UNK, ZSWIM3, ZNF335, MKRN2
SP_PIR_KEYWORDS	zinc	0,379	TRIM65, TRIM47, ZSWIM1, ZNF385D, RXRG, CHN2, RAF1, UNK, ZSWIM3, ZNF335, MKRN2
Annotation Cluster 9	Enrichment Score: 0.7169779381198176		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0000139~Golgi membrane	0,034	CAV2, CAV1, LARGE, DOPEY1
GOTERM_CC_FAT	GO:0044431~Golgi apparatus part	0,103	CAV2, CAV1, LARGE, DOPEY1
GOTERM_CC_FAT	GO:0005794~Golgi apparatus	0,241	CAV2, CAV1, LARGE, STIP1, DOPEY1, KIFC3
GOTERM_CC_FAT	GO:0031090~organelle membrane	0,241	CAV2, ACOX1, CAV1, LARGE, COX6B1, RAF1, DOPEY1
SP_PIR_KEYWORDS	golgi apparatus	0,420	CAV2, CAV1, LARGE, CALN1
GOTERM_CC_FAT	GO:0012505~endomembrane system	0,584	CAV2, CAV1, LARGE, DOPEY1
Annotation Cluster 10	Enrichment Score: 0.6890850699224897		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0016324~apical plasma membrane	0,092	CAV2, CAV1, ATP4A
GOTERM_BP_FAT	GO:0006811~ion transport	0,126	CAV2, CAV1, ATP4A, SLC22A23, CNGB1, SLC9A10
GOTERM_CC_FAT	GO:0045177~apical part of cell	0,150	CAV2, CAV1, ATP4A
GOTERM_BP_FAT	GO:0042493~response to drug	0,173	CAV2, CAV1, ATP4A
GOTERM_CC_FAT	GO:0005887~integral to plasma membrane	0,483	CAV2, FLRT1, CAV1, ATP4A, NLGN1, NRXN1
GOTERM_CC_FAT	GO:0031226~intrinsic to plasma membrane	0,504	CAV2, FLRT1, CAV1, ATP4A, NLGN1, NRXN1
Annotation Cluster 11	Enrichment Score: 0.6740982029934728		
Category	Term	PValue	Genes
INTERPRO	IPR000372:Leucine-rich repeat, cysteine-rich flanking region, N-terminal	0,059	FLRT1, LRRC52, LGR4
SMART	SM00013:LRRNT	0,070	FLRT1, LRRC52, LGR4
INTERPRO	IPR001611:Leucine-rich repeat	0,207	FLRT1, LRRC52, LGR4
UP_SEQ_FEATURE	repeat:LRR 5	0,234	FLRT1, LRRC52, LGR4
UP_SEQ_FEATURE	repeat:LRR 4	0,265	FLRT1, LRRC52, LGR4
UP_SEQ_FEATURE	repeat:LRR 3	0,334	FLRT1, LRRC52, LGR4
UP_SEQ_FEATURE	repeat:LRR 1	0,362	FLRT1, LRRC52, LGR4
UP_SEQ_FEATURE	repeat:LRR 2	0,363	FLRT1, LRRC52, LGR4
SP_PIR_KEYWORDS	leucine-rich repeat	0,370	FLRT1, LRRC52, LGR4
Annotation Cluster 12	Enrichment Score: 0.6507564364023363		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0032504~multicellular organism reproduction	0,088	GAPDHS, CAV2, ACOX1, CAV1, SLC9A10
GOTERM_BP_FAT	GO:0048609~reproductive process in a multicellular organism	0,088	GAPDHS, CAV2, ACOX1, CAV1, SLC9A10
GOTERM_BP_FAT	GO:0007283~spermatogenesis	0,290	GAPDHS, ACOX1, SLC9A10
GOTERM_BP_FAT	GO:0048232~male gamete generation	0,290	GAPDHS, ACOX1, SLC9A10
GOTERM_BP_FAT	GO:0007276~gamete generation	0,401	GAPDHS, ACOX1, SLC9A10
GOTERM_BP_FAT	GO:0019953~sexual reproduction	0,476	GAPDHS, ACOX1, SLC9A10
Annotation Cluster 13	Enrichment Score: 0.6476017723747797		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0007017~microtubule-based process	0,002	CAV2, CAV1, HAUS5, TUBB2B, UBE2C, KIFC3
SP_PIR_KEYWORDS	microtubule	0,236	HAUS5, TUBB2B, KIFC3
GOTERM_CC_FAT	GO:0005874~microtubule	0,283	HAUS5, TUBB2B, KIFC3
GOTERM_CC_FAT	GO:0044430~cytoskeletal part	0,300	HAUS5, TUBB2B, TNNC2, FERMT3, NLGN1, KIFC3
GOTERM_CC_FAT	GO:0005856~cytoskeleton	0,622	HAUS5, TUBB2B, TNNC2, FERMT3, NLGN1, KIFC3
GOTERM_CC_FAT	GO:0015630~microtubule cytoskeleton	0,628	HAUS5, TUBB2B, KIFC3

GOTERM_CC_FAT	GO:0043232~intracellular non-membrane-bounded organelle	0,688	HAUS5, TUBB2B, TNNC2, SMG6, FERMT3, NLGN1, MRPL38, RBPJ, KIFC3, RPA3
GOTERM_CC_FAT	GO:0043228~non-membrane-bounded organelle	0,688	HAUS5, TUBB2B, TNNC2, SMG6, FERMT3, NLGN1, MRPL38, RBPJ, KIFC3, RPA3
Annotation Cluster 14	Enrichment Score: 0.6317027812254767		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0016192~vesicle-mediated transport	0,049	CAV2, CAV1, UNC13D, NLGN1, LIN7C, DOPEY1
GOTERM_BP_FAT	GO:0008104~protein localization	0,190	CAV1, SNX21, NLGN1, LIN7C, CTSA, DOPEY1
GOTERM_BP_FAT	GO:0015031~protein transport	0,272	SNX21, NLGN1, LIN7C, CTSA, DOPEY1
GOTERM_BP_FAT	GO:0045184~establishment of protein localization	0,277	SNX21, NLGN1, LIN7C, CTSA, DOPEY1
GOTERM_BP_FAT	GO:0046907~intracellular transport	0,401	SMG6, NLGN1, CTSA, DOPEY1
SP_PIR_KEYWORDS	protein transport	0,582	SNX21, LIN7C, DOPEY1
Annotation Cluster 15	Enrichment Score: 0.5947022492866074		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0007017~microtubule-based process	0,002	CAV2, CAV1, HAUS5, TUBB2B, UBE2C, KIFC3
GOTERM_BP_FAT	GO:0000280~nuclear division	0,041	HAUS5, TUBB2B, UBE2C, CDK3
GOTERM_BP_FAT	GO:0007067~mitosis	0,041	HAUS5, TUBB2B, UBE2C, CDK3
GOTERM_BP_FAT	GO:0000087~M phase of mitotic cell cycle	0,043	HAUS5, TUBB2B, UBE2C, CDK3
GOTERM_BP_FAT	GO:0048285~organelle fission	0,045	HAUS5, TUBB2B, UBE2C, CDK3
GOTERM_BP_FAT	GO:0000279~M phase	0,106	HAUS5, TUBB2B, UBE2C, CDK3
GOTERM_BP_FAT	GO:0000278~mitotic cell cycle	0,137	HAUS5, TUBB2B, UBE2C, CDK3
SP_PIR_KEYWORDS	mitosis	0,167	HAUS5, UBE2C, CDK3
GOTERM_BP_FAT	GO:0022403~cell cycle phase	0,173	HAUS5, TUBB2B, UBE2C, CDK3
GOTERM_BP_FAT	GO:0051301~cell division	0,273	HAUS5, UBE2C, CDK3
SP_PIR_KEYWORDS	cell division	0,285	HAUS5, UBE2C, CDK3
GOTERM_BP_FAT	GO:0022402~cell cycle process	0,313	HAUS5, TUBB2B, UBE2C, CDK3
GOTERM_MF_FAT	GO:0000166~nucleotide binding	0,479	GAPDHS, ACOX1, RBM42, TUBB2B, ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
GOTERM_BP_FAT	GO:0007049~cell cycle	0,511	HAUS5, TUBB2B, UBE2C, CDK3
SP_PIR_KEYWORDS	cell cycle	0,554	HAUS5, UBE2C, CDK3
GOTERM_MF_FAT	GO:0030554~adenyl nucleotide binding	0,557	ACOX1, ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
GOTERM_MF_FAT	GO:0001883~purine nucleoside binding	0,572	ACOX1, ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
GOTERM_MF_FAT	GO:0001882~nucleoside binding	0,579	ACOX1, ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
GOTERM_MF_FAT	GO:0017076~purine nucleotide binding	0,600	ACOX1, TUBB2B, ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
SP_PIR_KEYWORDS	nucleotide-binding	0,678	TUBB2B, ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
GOTERM_MF_FAT	GO:0032559~adenyl ribonucleotide binding	0,681	ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
GOTERM_MF_FAT	GO:0032553~ribonucleotide binding	0,710	TUBB2B, ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
GOTERM_MF_FAT	GO:0032555~purine ribonucleotide binding	0,710	TUBB2B, ATP4A, RAF1, CNGB1, UBE2C, KIFC3, CDK3
UP_SEQ_FEATURE	active site:Proton acceptor	0,742	ACOX1, RAF1, CDK3
SP_PIR_KEYWORDS	atp-binding	0,786	ATP4A, RAF1, UBE2C, KIFC3, CDK3
GOTERM_MF_FAT	GO:0005524~ATP binding	0,824	ATP4A, RAF1, UBE2C, KIFC3, CDK3
UP_SEQ_FEATURE	nucleotide phosphate-binding region:ATP	0,901	RAF1, KIFC3, CDK3
Annotation Cluster 16	Enrichment Score: 0.5437422454006914		
Category	Term	PValue	Genes
SP_PIR_KEYWORDS	zymogen	0,050	AOAH, TMPRSS7, CTSA, CPVL
GOTERM_MF_FAT	GO:0008236~serine-type peptidase activity	0,145	TMPRSS7, CTSA, CPVL
GOTERM_MF_FAT	GO:0017171~serine hydrolase activity	0,148	TMPRSS7, CTSA, CPVL
GOTERM_BP_FAT	GO:0006508~proteolysis	0,302	UBE2CBP, TMPRSS7, CTSA, NEURL2, UBE2C, CPVL
SP_PIR_KEYWORDS	Protease	0,581	TMPRSS7, CTSA, CPVL
SP_PIR_KEYWORDS	hydrolase	0,598	ACOT8, ATP4A, SMG6, AOAH, TMPRSS7, CTSA, CPVL
GOTERM_MF_FAT	GO:0070011~peptidase activity, acting on L-amino acid peptides	0,620	TMPRSS7, CTSA, CPVL
GOTERM_MF_FAT	GO:0008233~peptidase activity	0,644	TMPRSS7, CTSA, CPVL
Annotation Cluster 17	Enrichment Score: 0.5346971861270113		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0016192~vesicle-mediated transport	0,049	CAV2, CAV1, UNC13D, NLGN1, LIN7C, DOPEY1
GOTERM_CC_FAT	GO:0030054~cell junction	0,136	CAV2, CAV1, NLGN1, LIN7C, NEURL2
GOTERM_CC_FAT	GO:0005624~membrane fraction	0,376	CAV2, CAV1, LIN7C, CNGB1, DNAJC4
GOTERM_CC_FAT	GO:0005626~insoluble fraction	0,403	CAV2, CAV1, LIN7C, CNGB1, DNAJC4
GOTERM_BP_FAT	GO:0050877~neurological system process	0,413	CAV2, NLGN1, LIN7C, NRXN1, CNGB1, KIFC3
GOTERM_CC_FAT	GO:0000267~cell fraction	0,606	CAV2, CAV1, LIN7C, CNGB1, DNAJC4
GOTERM_BP_FAT	GO:0010033~response to organic substance	0,723	CAV2, CAV1, DNAJC4
Annotation Cluster 18	Enrichment Score: 0.5201993846701835		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0006811~ion transport	0,126	CAV2, CAV1, ATP4A, SLC22A23, CNGB1, SLC9A10
GOTERM_BP_FAT	GO:0055085~transmembrane transport	0,317	ATP4A, SLC22A23, CNGB1, SLC9A10
SP_PIR_KEYWORDS	ion transport	0,409	ATP4A, SLC22A23, CNGB1, SLC9A10

SP_PIR_KEYWORDS	transport	0,507	ATP4A, SLC22A23, SNX21, LIN7C, CNGB1, DOPEY1, SLC9A10, PLTP
Annotation Cluster 19	Enrichment Score: 0.5026979631586449		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0016192~vesicle-mediated transport	0,049	CAV2, CAV1, UNC13D, NLGN1, LIN7C, DOPEY1
GOTERM_BP_FAT	GO:0016044~membrane organization	0,383	CAV2, CAV1, UNC13D
GOTERM_CC_FAT	GO:0031410~cytoplasmic vesicle	0,713	CAV2, CAV1, UNC13D
GOTERM_CC_FAT	GO:0031982~vesicle	0,735	CAV2, CAV1, UNC13D
Annotation Cluster 20	Enrichment Score: 0.4840165452232292		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0006811~ion transport	0,126	CAV2, CAV1, ATP4A, SLC22A23, CNGB1, SLC9A10
GOTERM_BP_FAT	GO:0030001~metal ion transport	0,484	CAV1, ATP4A, SLC9A10
GOTERM_BP_FAT	GO:0006812~cation transport	0,578	CAV1, ATP4A, SLC9A10
Annotation Cluster 21	Enrichment Score: 0.4815738101041168		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0007268~synaptic transmission	0,085	CAV2, NLGN1, LIN7C, NRXN1
GOTERM_BP_FAT	GO:0019226~transmission of nerve impulse	0,121	CAV2, NLGN1, LIN7C, NRXN1
GOTERM_CC_FAT	GO:0030054~cell junction	0,136	CAV2, CAV1, NLGN1, LIN7C, NEURL2
SP_PIR_KEYWORDS	cell adhesion	0,239	FLRT1, FERMT3, NLGN1, NRXN1
GOTERM_CC_FAT	GO:0044456~synapse part	0,243	NLGN1, LIN7C, NRXN1
GOTERM_BP_FAT	GO:0007267~cell-cell signaling	0,347	CAV2, NLGN1, LIN7C, NRXN1
GOTERM_CC_FAT	GO:0045202~synapse	0,397	NLGN1, LIN7C, NRXN1
GOTERM_BP_FAT	GO:0050877~neurological system process	0,413	CAV2, NLGN1, LIN7C, NRXN1, CNGB1, KIFC3
GOTERM_BP_FAT	GO:0007155~cell adhesion	0,442	FLRT1, FERMT3, NLGN1, NRXN1
GOTERM_BP_FAT	GO:0022610~biological adhesion	0,443	FLRT1, FERMT3, NLGN1, NRXN1
GOTERM_CC_FAT	GO:0005887~integral to plasma membrane	0,483	CAV2, FLRT1, CAV1, ATP4A, NLGN1, NRXN1
GOTERM_CC_FAT	GO:0031226~intrinsic to plasma membrane	0,504	CAV2, FLRT1, CAV1, ATP4A, NLGN1, NRXN1
GOTERM_CC_FAT	GO:0005886~plasma membrane	0,613	CAV2, FLRT1, CAV1, ATP4A, FERMT3, NLGN1, LIN7C, RAF1, TMPRSS7, NRXN1, SLC9A10, LGR4, RGS3, UPK1A, NEURL2
GOTERM_CC_FAT	GO:0044459~plasma membrane part	0,626	CAV2, FLRT1, CAV1, ATP4A, FERMT3, NLGN1, LIN7C, NRXN1, NEURL2
UP_SEQ_FEATURE	topological domain:Cytoplasmic	0,711	CAV2, FLRT1, CAV1, ATP4A, NLGN1, TMPRSS7, NRXN1, CNGB1, LGR4, LARGE, UPK1A, LRRC52, CALN1
Annotation Cluster 22	Enrichment Score: 0.4013514014745415		
Category	Term	PValue	Genes
SP_PIR_KEYWORDS	zinc-finger	0,147	TRIM65, TRIM47, ZSWIM1, ZNF385D, RXRG, CHN2, RAF1, UNK, ZSWIM3, ZNF335, MKRN2
GOTERM_MF_FAT	GO:0008270~zinc ion binding	0,372	TRIM65, TRIM47, ZSWIM1, ZNF385D, RXRG, CHN2, RAF1, UNK, ZSWIM3, ZNF335, MKRN2
SP_PIR_KEYWORDS	zinc	0,379	TRIM65, TRIM47, ZSWIM1, ZNF385D, RXRG, CHN2, RAF1, UNK, ZSWIM3, ZNF335, MKRN2
SP_PIR_KEYWORDS	metal-binding	0,411	ATP4A, SMG6, RXRG, RAF1, NRXN1, ZSWIM3, ZNF335, MKRN2, TRIM47, TRIM65, ZSWIM1, ZNF385D, CHN2, UNK
GOTERM_MF_FAT	GO:0046914~transition metal ion binding	0,488	TRIM65, TRIM47, ZSWIM1, SMG6, ZNF385D, RXRG, CHN2, RAF1, UNK, ZSWIM3, ZNF335, MKRN2
GOTERM_MF_FAT	GO:0046872~metal ion binding	0,508	ATP4A, TNNC2, SMG6, RXRG, RAF1, NRXN1, SLC9A10, ZSWIM3, ZNF335, MKRN2, TRIM65, TRIM47, ZSWIM1, ZNF385D, CHN2, CALN1, UNK
GOTERM_MF_FAT	GO:0043169~cation binding	0,526	ATP4A, TNNC2, SMG6, RXRG, RAF1, NRXN1, SLC9A10, ZSWIM3, ZNF335, MKRN2, TRIM65, TRIM47, ZSWIM1, ZNF385D, CHN2, CALN1, UNK
GOTERM_MF_FAT	GO:0043167~ion binding	0,554	ATP4A, TNNC2, SMG6, RXRG, RAF1, NRXN1, SLC9A10, ZSWIM3, ZNF335, MKRN2, TRIM65, TRIM47, ZSWIM1, ZNF385D, CHN2, CALN1, UNK
Annotation Cluster 23	Enrichment Score: 0.355374621798503 54		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0044265~cellular macromolecule catabolic process	0,243	SMG6, UBE2CBP, NEURL2, UBE2C, RPA3
GOTERM_BP_FAT	GO:0009057~macromolecule catabolic process	0,287	SMG6, UBE2CBP, NEURL2, UBE2C, RPA3
GOTERM_BP_FAT	GO:0006508~proteolysis	0,302	UBE2CBP, TMPRSS7, CTSA, NEURL2, UBE2C, CPVL
SP_PIR_KEYWORDS	ubl conjugation pathway	0,334	UBE2CBP, NEURL2, UBE2C, MKRN2
GOTERM_BP_FAT	GO:0043632~modification-dependent macromolecule catabolic process	0,599	UBE2CBP, NEURL2, UBE2C
GOTERM_BP_FAT	GO:0019941~modification-dependent protein catabolic process	0,599	UBE2CBP, NEURL2, UBE2C
GOTERM_BP_FAT	GO:0051603~proteolysis involved in cellular protein catabolic process	0,623	UBE2CBP, NEURL2, UBE2C

GOTERM_BP_FAT	GO:0044257~cellular protein catabolic process	0,626	UBE2CBP, NEURL2, UBE2C
GOTERM_BP_FAT	GO:0030163~protein catabolic process	0,643	UBE2CBP, NEURL2, UBE2C
Annotation Cluster 24	Enrichment Score: 0.2614090869354313		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0010605~negative regulation of macromolecule metabolic process	0,473	CAV1, RFX3, RBPJ, UBE2C
GOTERM_BP_FAT	GO:0051172~negative regulation of nitrogen compound metabolic process	0,543	CAV1, RFX3, RBPJ
GOTERM_BP_FAT	GO:0031327~negative regulation of cellular biosynthetic process	0,586	CAV1, RFX3, RBPJ
GOTERM_BP_FAT	GO:0009890~negative regulation of biosynthetic process	0,598	CAV1, RFX3, RBPJ
Annotation Cluster 25	Enrichment Score: 0.20106634183701602		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0044265~cellular macromolecule catabolic process	0,243	SMG6, UBE2CBP, NEURL2, UBE2C, RPA3
GOTERM_BP_FAT	GO:0009057~macromolecule catabolic process	0,287	SMG6, UBE2CBP, NEURL2, UBE2C, RPA3
GOTERM_BP_FAT	GO:0006259~DNA metabolic process	0,529	SMG6, RBPJ, RPA3
GOTERM_MF_FAT	GO:0003677~DNA binding	0,801	SUPT3H, SMG6, ETV2, RXRG, RFX3, RBPJ, ZNF335, RPA3
GOTERM_CC_FAT	GO:0031981~nuclear lumen	0,822	SUPT3H, SMG6, RBPJ, UBE2C, RPA3
GOTERM_CC_FAT	GO:0031974~membrane-enclosed lumen	0,858	SUPT3H, SMG6, COX6B1, RBPJ, UBE2C, RPA3
GOTERM_CC_FAT	GO:0005654~nucleoplasm	0,861	SUPT3H, UBE2C, RPA3
GOTERM_CC_FAT	GO:0070013~intracellular organelle lumen	0,924	SUPT3H, SMG6, RBPJ, UBE2C, RPA3
GOTERM_CC_FAT	GO:0043233~organelle lumen	0,933	SUPT3H, SMG6, RBPJ, UBE2C, RPA3
Annotation Cluster 26	Enrichment Score: 0.14852030399602092		
Category	Term	PValue	Genes
UP_SEQ_FEATURE	transmembrane region	0,599	CAV2, CAV1, FLRT1, ATP4A, SLC22A23, NLGN1, TMEM40, C20ORF165, TMPRSS7, CNGB1, NRXN1, SLC9A10, LGR4, TMEM147, C3ORF52, LARGE, UPK1A, LRRC52, CALN1, DNAJC4
SP_PIR_KEYWORDS	transmembrane	0,636	CAV2, CAV1, FLRT1, ATP4A, SLC22A23, NLGN1, TMEM40, C20ORF165, TMPRSS7, CNGB1, NRXN1, SLC9A10, LGR4, TMEM147, C3ORF52, LARGE, UPK1A, LRRC52, CALN1, DNAJC4
GOTERM_CC_FAT	GO:0016021~integral to membrane	0,698	CAV2, CAV1, FLRT1, ATP4A, SLC22A23, NLGN1, TMEM40, C20ORF165, TMPRSS7, CNGB1, NRXN1, SLC9A10, LGR4, TMEM147, C3ORF52, LARGE, UPK1A, LRRC52, CALN1, DNAJC4
UP_SEQ_FEATURE	topological domain:Cytoplasmic	0,711	CAV2, FLRT1, CAV1, ATP4A, NLGN1, TMPRSS7, NRXN1, CNGB1, LGR4, LARGE, UPK1A, LRRC52, CALN1
SP_PIR_KEYWORDS	membrane	0,729	CAV2, CAV1, FLRT1, ATP4A, SLC22A23, NLGN1, TMEM40, LIN7C, C20ORF165, TMPRSS7, CNGB1, NRXN1, SLC9A10, LGR4, TMEM147, C3ORF52, UNC13D, LARGE, RGS3, UPK1A, LRRC52, CHN2, CALN1, DNAJC4
GOTERM_CC_FAT	GO:0031224~intrinsic to membrane	0,766	CAV2, CAV1, FLRT1, ATP4A, SLC22A23, NLGN1, TMEM40, C20ORF165, TMPRSS7, CNGB1, NRXN1, SLC9A10, LGR4, TMEM147, C3ORF52, LARGE, UPK1A, LRRC52, CALN1, DNAJC4
UP_SEQ_FEATURE	topological domain:Extracellular	0,864	FLRT1, UPK1A, NLGN1, LRRC52, TMPRSS7, CALN1, NRXN1, CNGB1, LGR4
Annotation Cluster 27	Enrichment Score: 0.11565887398816146		
Category	Term	PValue	Genes
GOTERM_MF_FAT	GO:0004672~protein kinase activity	0,674	CAV2, RAF1, CDK3
GOTERM_BP_FAT	GO:0006468~protein amino acid phosphorylation	0,681	CAV2, RAF1, CDK3
GOTERM_BP_FAT	GO:0016310~phosphorylation	0,775	CAV2, RAF1, CDK3
GOTERM_BP_FAT	GO:0006793~phosphorus metabolic process	0,861	CAV2, RAF1, CDK3
GOTERM_BP_FAT	GO:0006796~phosphate metabolic process	0,861	CAV2, RAF1, CDK3
Annotation Cluster 28	Enrichment Score: 0.08539839753573955		
Category	Term	PValue	Genes
SP_PIR_KEYWORDS	glycoprotein	0,680	FLRT1, ATP4A, SLC22A23, NLGN1, TMPRSS7, CTSA, NRXN1, CPVL, SLC9A10, LGR4, LARGE, AOAH, DMKN, UPK1A, LRRC52, PLTP, WFDC3
UP_SEQ_FEATURE	signal peptide	0,766	FLRT1, AOAH, DMKN, SBSN, NLGN1, LRRC52, CTSA, NRXN1, CPVL, LGR4, PLTP, WFDC3
SP_PIR_KEYWORDS	signal	0,774	FLRT1, AOAH, DMKN, SBSN, NLGN1, LRRC52, CTSA, NRXN1, CPVL, LGR4, PLTP, WFDC3
SP_PIR_KEYWORDS	disulfide bond	0,847	AOAH, UPK1A, COX6B1, NLGN1, TMPRSS7, CTSA, NRXN1, LGR4, PLTP, WFDC3
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	0,865	FLRT1, SLC22A23, NLGN1, CTSA, NRXN1, CPVL, SLC9A10,

			LGR4, LARGE, UPK1A, AOAH, LRRC52, PLTP, WFDC3
UP_SEQ_FEATURE	disulfide bond	0,890	AOAH, COX6B1, NLGN1, TMPRSS7, CTSA, NRXN1, LGR4, PLTP, WFDC3
GOTERM_CC_FAT	GO:0005576~extracellular region	0,961	FLRT1, DMKN, SBSN, PLTP, WFDC3
Annotation Cluster 29	Enrichment Score: 0.08503105669816337		
Category	Term	PValue	Genes
GOTERM_MF_FAT	GO:0043565~sequence-specific DNA binding	0,404	SMG6, ETV2, RXRG, RFX3
GOTERM_MF_FAT	GO:0003700~transcription factor activity	0,723	ETV2, RXRG, RFX3, RBPJ
GOTERM_MF_FAT	GO:0003677~DNA binding	0,801	SUPT3H, SMG6, ETV2, RXRG, RFX3, RBPJ, ZNF335, RPA3
GOTERM_MF_FAT	GO:0030528~transcription regulator activity	0,838	SUPT3H, ETV2, RXRG, RFX3, RBPJ
SP_PIR_KEYWORDS	dna-binding	0,880	SMG6, ETV2, RXRG, RFX3, RBPJ, ZNF335
GOTERM_BP_FAT	GO:0006355~regulation of transcription, DNA-dependent	0,881	SUPT3H, ETV2, RXRG, RFX3, RBPJ
GOTERM_BP_FAT	GO:0051252~regulation of RNA metabolic process	0,891	SUPT3H, ETV2, RXRG, RFX3, RBPJ
GOTERM_BP_FAT	GO:0006350~transcription	0,948	SUPT3H, RXRG, RFX3, RBPJ, ZNF335
GOTERM_BP_FAT	GO:0045449~regulation of transcription	0,962	SUPT3H, ETV2, RXRG, RFX3, RBPJ, ZNF335
SP_PIR_KEYWORDS	transcription regulation	0,968	SUPT3H, RXRG, RFX3, RBPJ, ZNF335
SP_PIR_KEYWORDS	Transcription	0,972	SUPT3H, RXRG, RFX3, RBPJ, ZNF335
Kumyks Adygei			
Annotation Cluster 1	Enrichment Score: 1.982791165248524		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0043413~biopolymer glycosylation	536.856.826.624,374	ST6GALNAC6, ST6GALNAC4, DPM2, TSPAN8, ALK
GOTERM_BP_FAT	GO:0070085~glycosylation	536.856.826.624,374	ST6GALNAC6, ST6GALNAC4, DPM2, TSPAN8, ALK
GOTERM_BP_FAT	GO:0006486~protein amino acid glycosylation	536.856.826.624,374	ST6GALNAC6, ST6GALNAC4, DPM2, TSPAN8, ALK
GOTERM_BP_FAT	GO:0009101~glycoprotein biosynthetic process	0,001	ST6GALNAC6, ST6GALNAC4, DPM2, TSPAN8, ALK
GOTERM_BP_FAT	GO:0009100~glycoprotein metabolic process	0,003	ST6GALNAC6, ST6GALNAC4, DPM2, TSPAN8, ALK
SP_PIR_KEYWORDS	transferase	0,040	ST6GALNAC6, PAK7, ST6GALNAC4, PIP5KL1, ERBB4, AK1, NEK1, DPM2, ALK
GOTERM_CC_FAT	GO:0012505~endomembrane system	0,041	ST6GALNAC6, ST6GALNAC4, DENND1A, MAPK8IP3, DPM2, XPO7, RAB27A
GOTERM_CC_FAT	GO:0031301~integral to organelle membrane	0,062	ST6GALNAC6, ST6GALNAC4, DPM2
GOTERM_CC_FAT	GO:0031300~intrinsic to organelle membrane	0,083	ST6GALNAC6, ST6GALNAC4, DPM2
SP_PIR_KEYWORDS	glycosyltransferase	0,121	ST6GALNAC6, ST6GALNAC4, DPM2
GOTERM_CC_FAT	GO:0031090~organelle membrane	0,290	ST6GALNAC6, ST6GALNAC4, DENND1A, MAPK8IP3, DPM2, RAB27A
Annotation Cluster 2	Enrichment Score: 1.7159439024947676		
Category	Term	PValue	Genes
UP_SEQ_FEATURE	short sequence motif:SLAM-like motif	77.923.696.851,306	SIGLEC6, SIGLEC5, SIGLEC12
UP_SEQ_FEATURE	binding site:Sialic acid	599.051.003.861,574	SIGLEC6, SIGLEC5, SIGLEC14
UP_SEQ_FEATURE	short sequence motif:ITIM motif	803.498.184.851,187	SIGLEC6, SIGLEC5, SIGLEC12
GOTERM_MF_FAT	GO:0005529~sugar binding	0,002	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14, ENG
GOTERM_MF_FAT	GO:0030246~carbohydrate binding	0,003	SIGLEC6, COL13A1, SIGLEC5, SIGLEC12, SIGLEC14, ENG
SP_PIR_KEYWORDS	cell adhesion	0,006	SIGLEC6, COL13A1, SIGLEC5, SIGLEC12, SIGLEC14, ENG
SP_PIR_KEYWORDS	Lectin	0,010	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
UP_SEQ_FEATURE	domain:Ig-like C2-type 1	0,017	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
UP_SEQ_FEATURE	domain:Ig-like C2-type 2	0,017	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
INTERPRO	IPR013151:Immunoglobulin	0,022	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
INTERPRO	IPR003598:Immunoglobulin subtype 2	0,022	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
INTERPRO	IPR003006:Immunoglobulin/major histocompatibility complex, conserved site	0,027	SIGLEC6, SIGLEC5, SIGLEC14
SMART	SM00408:IGc2	0,029	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
INTERPRO	IPR013106:Immunoglobulin V-set	0,037	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
UP_SEQ_FEATURE	domain:Ig-like V-type	0,048	SIGLEC6, SIGLEC5, SIGLEC14
GOTERM_BP_FAT	GO:0007155~cell adhesion	0,054	SIGLEC6, COL13A1, SIGLEC5, SIGLEC12, SIGLEC14, ENG
GOTERM_BP_FAT	GO:0022610~biological adhesion	0,054	SIGLEC6, COL13A1, SIGLEC5, SIGLEC12, SIGLEC14, ENG
INTERPRO	IPR003599:Immunoglobulin subtype	0,073	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
SMART	SM00409:IG	0,094	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14

SP_PIR_KEYWORDS	Immunoglobulin domain	0,145	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
INTERPRO	IPR007110:Immunoglobulin-like	0,182	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
INTERPRO	IPR013783:Immunoglobulin-like fold	0,222	SIGLEC6, SIGLEC5, SIGLEC12, SIGLEC14
GOTERM_CC_FAT	GO:0005576~extracellular region	0,914	NHLRC3, SIGLEC6, COL13A1, FBN2, ENG
Annotation Cluster 3	Enrichment Score: 1.3771645812669877		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0046328~regulation of JNK cascade	0,016	SH3RF1, GRIK2, MAPK8IP3
GOTERM_BP_FAT	GO:0070302~regulation of stress-activated protein kinase signaling pathway	0,018	SH3RF1, GRIK2, MAPK8IP3
GOTERM_BP_FAT	GO:0080135~regulation of cellular response to stress	0,037	SH3RF1, GRIK2, MAPK8IP3
GOTERM_BP_FAT	GO:0043408~regulation of MAPKKK cascade	0,041	SH3RF1, GRIK2, MAPK8IP3
GOTERM_CC_FAT	GO:0042995~cell projection	0,077	SH3RF1, PIP5KL1, GRIK2, DENND1A, MAPK8IP3, RAB27A
GOTERM_BP_FAT	GO:0010627~regulation of protein kinase cascade	0,167	SH3RF1, GRIK2, MAPK8IP3
Annotation Cluster 4	Enrichment Score: 1.2861031299951242		
Category	Term	PValue	Genes
SP_PIR_KEYWORDS	glycoprotein	0,011	TMEM204, ERBB4, COL13A1, GRIK2, TSPAN8, SIGLEC12, ALK, SIGLEC14, NAALADL2, CSMD1, ST6GALNAC6, TMEM117, ST6GALNAC4, SIGLEC6, SEMA6D, NHLRC3, SIGLEC5, C7ORF58, FBN2, ENG, GFRA2
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	0,015	TMEM204, ERBB4, GRIK2, TSPAN8, SIGLEC12, ALK, SIGLEC14, NAALADL2, CSMD1, ST6GALNAC6, TMEM117, ST6GALNAC4, SIGLEC6, SEMA6D, NHLRC3, SIGLEC5, C7ORF58, FBN2, ENG, GFRA2
SP_PIR_KEYWORDS	membrane	0,024	TM7SF4, ERBB4, GRIK2, TSPAN8, RIMS2, NAALADL2, CSMD1, ST6GALNAC6, ST6GALNAC4, PIP5KL1, RAB27A, TMEM204, COL13A1, DENND1A, SIGLEC12, ALK, SIGLEC14, TMEM117, SIGLEC6, SEMA6D, PLEKHH2, SIGLEC5, DPM2, SLC27A6, ENG, GFRA2
SP_PIR_KEYWORDS	transmembrane	0,024	TM7SF4, TMEM204, ERBB4, COL13A1, GRIK2, TSPAN8, SIGLEC12, ALK, SIGLEC14, NAALADL2, CSMD1, ST6GALNAC6, TMEM117, ST6GALNAC4, SIGLEC6, SEMA6D, PLEKHH2, SIGLEC5, SLC27A6, DPM2, ENG, IFT140
UP_SEQ_FEATURE	topological domain:Extracellular	0,037	TMEM204, ERBB4, GRIK2, COL13A1, TSPAN8, SIGLEC12, ALK, SIGLEC14, NAALADL2, CSMD1, SEMA6D, SIGLEC6, SIGLEC5, ENG
UP_SEQ_FEATURE	transmembrane region	0,043	TM7SF4, TMEM204, ERBB4, COL13A1, GRIK2, TSPAN8, SIGLEC12, ALK, SIGLEC14, NAALADL2, CSMD1, ST6GALNAC6, TMEM117, ST6GALNAC4, SEMA6D, SIGLEC6, PLEKHH2, SIGLEC5, SLC27A6, DPM2, ENG
UP_SEQ_FEATURE	topological domain:Cytoplasmic	0,043	TM7SF4, TMEM204, ERBB4, GRIK2, COL13A1, TSPAN8, SIGLEC12, ALK, NAALADL2, CSMD1, ST6GALNAC6, ST6GALNAC4, SEMA6D, SIGLEC6, SIGLEC5, ENG
GOTERM_CC_FAT	GO:0031224~intrinsic to membrane	0,044	TM7SF4, ERBB4, GRIK2, TSPAN8, NAALADL2, CSMD1, ST6GALNAC6, ST6GALNAC4, RAB27A, IFT140, TMEM204, COL13A1, SIGLEC12, ALK, SIGLEC14, TMEM117, SIGLEC6, SEMA6D, PLEKHH2, SIGLEC5, DPM2, SLC27A6, XPO7, ENG, GFRA2
GOTERM_CC_FAT	GO:0016021~integral to membrane	0,100	TM7SF4, TMEM204, ERBB4, COL13A1, GRIK2, TSPAN8, SIGLEC12, ALK, SIGLEC14, NAALADL2, CSMD1, ST6GALNAC6, TMEM117, ST6GALNAC4, SIGLEC6, SEMA6D, PLEKHH2, SIGLEC5, SLC27A6, DPM2, XPO7, ENG, IFT140
SP_PIR_KEYWORDS	signal	0,113	ERBB4, GRIK2, SIGLEC12, ALK, SIGLEC14, CSMD1, NHLRC3, SEMA6D, SIGLEC6, SIGLEC5, C7ORF58, FBN2, ENG, GFRA2
UP_SEQ_FEATURE	signal peptide	0,117	ERBB4, GRIK2, SIGLEC12, ALK, SIGLEC14, CSMD1, NHLRC3, SEMA6D, SIGLEC6, SIGLEC5, C7ORF58, FBN2, ENG, GFRA2
SP_PIR_KEYWORDS	disulfide bond	0,189	ST6GALNAC6, ST6GALNAC4, ERBB4, SIGLEC6, SEMA6D, COL13A1, SIGLEC5, SIGLEC12, FBN2, SIGLEC14, RAB27A, CSMD1
UP_SEQ_FEATURE	disulfide bond	0,268	ST6GALNAC6, ST6GALNAC4, ERBB4, SIGLEC6, SEMA6D, SIGLEC5, SIGLEC12, FBN2, SIGLEC14, RAB27A, CSMD1
Annotation Cluster 5	Enrichment Score: 1.2014658239787752		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0030384~phosphoinositide metabolic process	0,020	PIP5KL1, SEMA6D, DPM2
GOTERM_BP_FAT	GO:0006650~glycerophospholipid metabolic	0,047	PIP5KL1, SEMA6D, DPM2

	process		
GOTERM_BP_FAT	GO:0046486~glycerolipid metabolic process	0,083	PIP5KL1, SEMA6D, DPM2
GOTERM_BP_FAT	GO:0006644~phospholipid metabolic process	0,108	PIP5KL1, SEMA6D, DPM2
GOTERM_BP_FAT	GO:0019637~organophosphate metabolic process	0,118	PIP5KL1, SEMA6D, DPM2
Annotation Cluster 6	Enrichment Score: 1.0785741367117818		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0006886~intracellular protein transport	0,024	GRIK2, RIMS2, XPO7, RAB27A, KIF13B
GOTERM_BP_FAT	GO:0034613~cellular protein localization	0,032	GRIK2, RIMS2, XPO7, RAB27A, KIF13B
GOTERM_BP_FAT	GO:0070727~cellular macromolecule localization	0,033	GRIK2, RIMS2, XPO7, RAB27A, KIF13B
GOTERM_BP_FAT	GO:0008104~protein localization	0,043	GRIK2, MAPK8IP3, RIMS2, XPO7, RAB27A, EXOC6B, KIF13B
GOTERM_BP_FAT	GO:0015031~protein transport	0,072	GRIK2, RIMS2, XPO7, RAB27A, EXOC6B, KIF13B
GOTERM_BP_FAT	GO:0045184~establishment of protein localization	0,075	GRIK2, RIMS2, XPO7, RAB27A, EXOC6B, KIF13B
GOTERM_BP_FAT	GO:0046907~intracellular transport	0,128	GRIK2, RIMS2, XPO7, RAB27A, KIF13B
GOTERM_BP_FAT	GO:0006605~protein targeting	0,133	XPO7, RAB27A, KIF13B
GOTERM_MF_FAT	GO:0019899~enzyme binding	0,186	GRIK2, MAPK8IP3, RIMS2, KIF13B
SP_PIR_KEYWORDS	transport	0,859	GRIK2, SLC27A6, XPO7, EXOC6B
Annotation Cluster 7	Enrichment Score: 1.064483004777457		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0030425~dendrite	0,016	GRIK2, DENND1A, MAPK8IP3, RAB27A
GOTERM_CC_FAT	GO:0012505~endomembrane system	0,041	ST6GALNAC6, ST6GALNAC4, DENND1A, MAPK8IP3, DPM2, XPO7, RAB27A
GOTERM_CC_FAT	GO:0043005~neuron projection	0,102	GRIK2, DENND1A, MAPK8IP3, RAB27A
GOTERM_BP_FAT	GO:0016192~vesicle-mediated transport	0,242	DENND1A, MAPK8IP3, RAB27A, EXOC6B
GOTERM_CC_FAT	GO:0031090~organelle membrane	0,290	ST6GALNAC6, ST6GALNAC4, DENND1A, MAPK8IP3, DPM2, RAB27A
Annotation Cluster 8	Enrichment Score: 1.0501339780942553		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0030054~cell junction	0,027	TMEM204, GRIK2, COL13A1, DENND1A, RIMS2, IFT140
SP_PIR_KEYWORDS	cell junction	0,102	TMEM204, GRIK2, DENND1A, RIMS2
GOTERM_CC_FAT	GO:0045202~synapse	0,111	ERBB4, GRIK2, DENND1A, RIMS2
SP_PIR_KEYWORDS	synapse	0,120	GRIK2, DENND1A, RIMS2
SP_PIR_KEYWORDS	cell membrane	0,157	TM7SF4, TMEM204, SIGLEC6, SEMA6D, GRIK2, COL13A1, DENND1A, SLC27A6, RIMS2, GFRA2
Annotation Cluster 9	Enrichment Score: 1.040952272281866		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0007167~enzyme linked receptor protein signaling pathway	0,003	DOK2, ERBB4, GRIK2, ALK, ENG, GFRA2
GOTERM_BP_FAT	GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	0,004	DOK2, ERBB4, GRIK2, ALK, GFRA2
GOTERM_BP_FAT	GO:0010647~positive regulation of cell communication	0,254	ERBB4, GRIK2, ENG
SP_PIR_KEYWORDS	receptor	0,461	ERBB4, GRIK2, RORA, ALK, THADA, GFRA2
GOTERM_MF_FAT	GO:0042802~identical protein binding	0,551	DOK2, GRIK2, ENG
GOTERM_BP_FAT	GO:0007166~cell surface receptor linked signal transduction	0,658	DOK2, ERBB4, GRIK2, ALK, ENG, GFRA2
Annotation Cluster 10	Enrichment Score: 0.9160849156813619		
Category	Term	PValue	Genes
GOTERM_MF_FAT	GO:0004713~protein tyrosine kinase activity	0,012	TWF1, ERBB4, NEK1, ALK
GOTERM_MF_FAT	GO:0004672~protein kinase activity	0,028	PAK7, TWF1, ERBB4, NEK1, ALK, ENG
SP_PIR_KEYWORDS	tyrosine-protein kinase	0,039	ERBB4, NEK1, ALK
SP_PIR_KEYWORDS	transferase	0,040	ST6GALNAC6, PAK7, ST6GALNAC4, PIP5KL1, ERBB4, AK1, NEK1, DPM2, ALK
SP_PIR_KEYWORDS	kinase	0,043	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK
GOTERM_BP_FAT	GO:0006468~protein amino acid phosphorylation	0,045	PAK7, TWF1, ERBB4, NEK1, MAPK8IP3, ALK
GOTERM_BP_FAT	GO:0006793~phosphorus metabolic process	0,064	PAK7, DUSP4, TWF1, ERBB4, NEK1, MAPK8IP3, ALK
GOTERM_BP_FAT	GO:0006796~phosphate metabolic process	0,064	PAK7, DUSP4, TWF1, ERBB4, NEK1, MAPK8IP3, ALK
GOTERM_BP_FAT	GO:0016310~phosphorylation	0,085	PAK7, TWF1, ERBB4, NEK1, MAPK8IP3, ALK
SP_PIR_KEYWORDS	nucleotide-binding	0,097	PAK7, PIP5KL1, ERBB4, AK1, NEK1, SLC27A6, ALK, RAB27A, KIF13B
UP_SEQ_FEATURE	nucleotide phosphate-binding region:ATP	0,134	PAK7, ERBB4, AK1, NEK1, ALK, KIF13B
UP_SEQ_FEATURE	domain:Protein kinase	0,146	PAK7, ERBB4, NEK1, ALK
INTERPRO	IPR017441:Protein kinase, ATP binding site	0,149	PAK7, ERBB4, NEK1, ALK
INTERPRO	IPR000719:Protein kinase, core	0,164	PAK7, ERBB4, NEK1, ALK
SP_PIR_KEYWORDS	atp-binding	0,166	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, KIF13B
UP_SEQ_FEATURE	binding site:ATP	0,197	PAK7, ERBB4, NEK1, ALK

GOTERM_MF_FAT	GO:0005524~ATP binding	0,240	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, KIF13B
GOTERM_MF_FAT	GO:0032559~adenyl ribonucleotide binding	0,250	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, KIF13B
GOTERM_MF_FAT	GO:0032555~purine ribonucleotide binding	0,263	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, RAB27A, KIF13B
GOTERM_MF_FAT	GO:0032553~ribonucleotide binding	0,263	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, RAB27A, KIF13B
UP_SEQ_FEATURE	active site:Proton acceptor	0,284	PAK7, ERBB4, NEK1, ALK
GOTERM_MF_FAT	GO:0030554~adenyl nucleotide binding	0,290	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, KIF13B
GOTERM_MF_FAT	GO:0017076~purine nucleotide binding	0,301	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, RAB27A, KIF13B
GOTERM_MF_FAT	GO:0001883~purine nucleoside binding	0,302	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, KIF13B
GOTERM_MF_FAT	GO:0000166~nucleotide binding	0,303	PAK7, PIP5KL1, ERBB4, AK1, NEK1, SLC27A6, ALK, RAB27A, KIF13B
GOTERM_MF_FAT	GO:0001882~nucleoside binding	0,308	PAK7, PIP5KL1, ERBB4, AK1, NEK1, ALK, KIF13B
Annotation Cluster 11	Enrichment Score: 0.7842069029929679		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0012505~endomembrane system	0,041	ST6GALNAC6, ST6GALNAC4, DENND1A, MAPK8IP3, DPM2, XPO7, RAB27A
GOTERM_CC_FAT	GO:0000139~Golgi membrane	0,124	ST6GALNAC6, ST6GALNAC4, MAPK8IP3
GOTERM_CC_FAT	GO:0044431~Golgi apparatus part	0,252	ST6GALNAC6, ST6GALNAC4, MAPK8IP3
GOTERM_CC_FAT	GO:0031090~organelle membrane	0,290	ST6GALNAC6, ST6GALNAC4, DENND1A, MAPK8IP3, DPM2, RAB27A
GOTERM_CC_FAT	GO:0005794~Golgi apparatus	0,321	ST6GALNAC6, SH3RF1, ST6GALNAC4, MAPK8IP3, RAB27A
Annotation Cluster 12	Enrichment Score: 0.7364615260002166		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0042995~cell projection	0,077	SH3RF1, PIP5KL1, GRIK2, DENND1A, MAPK8IP3, RAB27A
GOTERM_BP_FAT	GO:0043068~positive regulation of programmed cell death	0,135	SH3RF1, ING3, GRIK2, RAB27A
GOTERM_BP_FAT	GO:0010942~positive regulation of cell death	0,137	SH3RF1, ING3, GRIK2, RAB27A
GOTERM_BP_FAT	GO:0042981~regulation of apoptosis	0,212	PAK7, SH3RF1, ING3, GRIK2, RAB27A
GOTERM_BP_FAT	GO:0043067~regulation of programmed cell death	0,217	PAK7, SH3RF1, ING3, GRIK2, RAB27A
GOTERM_BP_FAT	GO:0010941~regulation of cell death	0,219	PAK7, SH3RF1, ING3, GRIK2, RAB27A
GOTERM_BP_FAT	GO:0012502~induction of programmed cell death	0,245	SH3RF1, GRIK2, RAB27A
GOTERM_BP_FAT	GO:0043065~positive regulation of apoptosis	0,365	SH3RF1, ING3, RAB27A
Annotation Cluster 13	Enrichment Score: 0.6065620670043955		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0000904~cell morphogenesis involved in differentiation	0,162	LHX2, MAPK8IP3, PROX1
GOTERM_BP_FAT	GO:0000902~cell morphogenesis	0,284	LHX2, MAPK8IP3, PROX1
GOTERM_BP_FAT	GO:0032989~cellular component morphogenesis	0,329	LHX2, MAPK8IP3, PROX1
Annotation Cluster 14	Enrichment Score: 0.2840752009665476		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0070271~protein complex biogenesis	0,443	GRIK2, DPYS, XPO7
GOTERM_BP_FAT	GO:0006461~protein complex assembly	0,443	GRIK2, DPYS, XPO7
GOTERM_BP_FAT	GO:0065003~macromolecular complex assembly	0,592	GRIK2, DPYS, XPO7
GOTERM_BP_FAT	GO:0043933~macromolecular complex subunit organization	0,628	GRIK2, DPYS, XPO7
Annotation Cluster 15	Enrichment Score: 0.20268871611526484		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0000267~cell fraction	0,282	ST6GALNAC4, SIGLEC6, GRIK2, MAPK8IP3, DPYS, ENG
GOTERM_CC_FAT	GO:0005624~membrane fraction	0,754	SIGLEC6, GRIK2, ENG
GOTERM_CC_FAT	GO:0005887~integral to plasma membrane	0,762	SIGLEC6, GRIK2, ALK, ENG
GOTERM_CC_FAT	GO:0005626~insoluble fraction	0,772	SIGLEC6, GRIK2, ENG
GOTERM_CC_FAT	GO:0031226~intrinsic to plasma membrane	0,776	SIGLEC6, GRIK2, ALK, ENG
Annotation Cluster 16	Enrichment Score: 0.1629444103158404		
Category	Term	PValue	Genes
GOTERM_BP_FAT	GO:0045893~positive regulation of transcription, DNA-dependent	0,414	RORA, PROX1, ENG
GOTERM_BP_FAT	GO:0051254~positive regulation of RNA metabolic process	0,419	RORA, PROX1, ENG
GOTERM_BP_FAT	GO:0045941~positive regulation of transcription	0,501	RORA, PROX1, ENG
GOTERM_BP_FAT	GO:0010628~positive regulation of gene expression	0,517	RORA, PROX1, ENG
GOTERM_MF_FAT	GO:0043565~sequence-specific DNA binding	0,522	LHX2, RORA, PROX1
GOTERM_MF_FAT	GO:0003700~transcription factor activity	0,533	LHX2, RORA, PROX1, ZNF175
GOTERM_BP_FAT	GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0,557	RORA, PROX1, ENG
GOTERM_BP_FAT	GO:0051173~positive regulation of nitrogen	0,574	RORA, PROX1, ENG

	compound metabolic process		
GOTERM_BP_FAT	GO:0010557~positive regulation of macromolecule biosynthetic process	0,583	RORA, PROX1, ENG
SP_PIR_KEYWORDS	zinc	0,589	SH3RF1, ING3, LHX2, DPYS, RORA, RIMS2, ZNF175
GOTERM_BP_FAT	GO:0031328~positive regulation of cellular biosynthetic process	0,608	RORA, PROX1, ENG
GOTERM_BP_FAT	GO:0009891~positive regulation of biosynthetic process	0,616	RORA, PROX1, ENG
GOTERM_BP_FAT	GO:0006357~regulation of transcription from RNA polymerase II promoter	0,641	RORA, PROX1, ENG
GOTERM_MF_FAT	GO:0008270~zinc ion binding	0,666	SH3RF1, ING3, LHX2, DPYS, RORA, RIMS2, ZNF175
SP_PIR_KEYWORDS	nucleus	0,683	DUSP4, ING3, LHX2, NEK1, RAPGEF5, RSL24D1, RORA, CRAMP1L, XPO7, PROX1, ZNF175, HN1L
SP_PIR_KEYWORDS	zinc-finger	0,722	SH3RF1, ING3, RORA, RIMS2, ZNF175
GOTERM_BP_FAT	GO:0010604~positive regulation of macromolecule metabolic process	0,730	RORA, PROX1, ENG
SP_PIR_KEYWORDS	metal-binding	0,749	SH3RF1, ING3, LHX2, NEK1, DPYS, RORA, RIMS2, ZNF175
SP_PIR_KEYWORDS	dna-binding	0,782	LHX2, RORA, CRAMP1L, PROX1, ZNF175
GOTERM_BP_FAT	GO:0006355~regulation of transcription, DNA-dependent	0,788	LHX2, RORA, PROX1, ENG, ZNF175
GOTERM_BP_FAT	GO:0051252~regulation of RNA metabolic process	0,803	LHX2, RORA, PROX1, ENG, ZNF175
GOTERM_MF_FAT	GO:0030528~transcription regulator activity	0,822	LHX2, RORA, PROX1, ZNF175
SP_PIR_KEYWORDS	transcription regulation	0,834	ING3, LHX2, RORA, PROX1, ZNF175
GOTERM_MF_FAT	GO:0046914~transition metal ion binding	0,836	SH3RF1, ING3, LHX2, DPYS, RORA, RIMS2, ZNF175
SP_PIR_KEYWORDS	Transcription	0,847	ING3, LHX2, RORA, PROX1, ZNF175
GOTERM_BP_FAT	GO:0006350~transcription	0,888	ING3, LHX2, RORA, PROX1, ZNF175
GOTERM_BP_FAT	GO:0045449~regulation of transcription	0,906	ING3, LHX2, RORA, PROX1, ENG, ZNF175
GOTERM_MF_FAT	GO:0003677~DNA binding	0,919	LHX2, RORA, CRAMP1L, PROX1, ZNF175
GOTERM_MF_FAT	GO:0046872~metal ion binding	0,940	SH3RF1, ING3, LHX2, NEK1, DPYS, RORA, FBN2, RIMS2, ZNF175
GOTERM_MF_FAT	GO:0043169~cation binding	0,944	SH3RF1, ING3, LHX2, NEK1, DPYS, RORA, FBN2, RIMS2, ZNF175
GOTERM_MF_FAT	GO:0043167~ion binding	0,951	SH3RF1, ING3, LHX2, NEK1, DPYS, RORA, FBN2, RIMS2, ZNF175
Annotation Cluster 17	Enrichment Score: 0.07162425132942471		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0005856~cytoskeleton	0,679	TWF1, ERBB4, PLEKHH2, ZNF175, KIF13B
GOTERM_CC_FAT	GO:0043232~intracellular non-membrane-bounded organelle	0,948	TWF1, ERBB4, PLEKHH2, RSL24D1, ZNF175, KIF13B
GOTERM_CC_FAT	GO:0043228~non-membrane-bounded organelle	0,948	TWF1, ERBB4, PLEKHH2, RSL24D1, ZNF175, KIF13B
Annotation Cluster 18	Enrichment Score: 0.008606696963561034		
Category	Term	PValue	Genes
GOTERM_CC_FAT	GO:0031981~nuclear lumen	0,960	DUSP4, ING3, RSL24D1
GOTERM_CC_FAT	GO:0070013~intracellular organelle lumen	0,986	DUSP4, ING3, RSL24D1
GOTERM_CC_FAT	GO:0043233~organelle lumen	0,988	DUSP4, ING3, RSL24D1
GOTERM_CC_FAT	GO:0031974~membrane-enclosed lumen	0,989	DUSP4, ING3, RSL24D1