CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty ot Tropical AgriSciences

Department of Animal Science and Food Processing

Czech University of Life Sciences Prague
Faculty of Tropical AgriSciences

The origin and evolution of functional haemoglobin polymorphism in the bank vole (*Clethrionomys glareolus*)

Master thesis

Prague 2015

Supervisor:
RNDr. Petr Kotlík, Ph.D.
Institute of Animal Physiology and Genetics CAS, v. v. i.

Consultant:
RNDr. Silvia Marková, Ph.D.
Institute of Animal Physiology and Genetics CAS, v. v. i.

Author:
Bc. Michaela Strážnická

Declaration:

I hereby declare, that this thesis entitled 'The origin and evolution of functional haemoglobin polymorphism in the bank vole (*Clethrionomys glareolus*)' is my own work. All sources of information have been quoted and acknowledged by means of complete references. No part of this thesis have been used to acquire same or other academic title.

In Prague, 22.4. 2015

Michaela Strážnická

Acknowledgements:

My greatest thanks go to my supervisor RNDr. Petr Kotlík, Ph.D. for his trust in my abilities, for constructive discussions and also for his invaluable help and endless patience when this thesis was becoming reality. Next, I would like to thank my thesis consultant RNDr. Silvia Marková, Ph.D., who introduced me into the magic world of the laboratory, for her eager help with simply any problem. My thanks go further to our great technician Bc. Jana Kopecká, who maintained our material and chemical stock always full and ready and who helped us with laborous procedures. Also I would like to thank to Bc. Karolína Filipi for her constructive and friendly talks and also for few pictures she allowed me to borrow from her thesis.

My next great thanks goes to members of our faculty, namely Ing. Karolína Brandlová, Ph.D. and doc. Ing. Jan Banout, Ph.D. who were really helpful in making this interinstitutional cooperation possible.

Really huge acknowledgement goes to my family for their support during my whole studies. I would like to thank especially to my mother, who enabled me to study abroad for one year during which my interest in molecular methods was formed and who supported me indulgently all the way long to this point. Even though it might not be evident, I am aware of the fact it has not been always a piece of cake and I really appreciate it.

Last but not least I would like to thank to my friends for their support. A really big hug and kiss deserves namely my boyfriend Ondra who appreciates my work even though it is all Greek to him.

My work was financially supported by the Czech Science Foundation (grant number P506-11-1872) and by the institutional support of IAPG CAS, v. v. i. (RVO 67985904). I also received an intra-institutional grant from the IAPG CAS, v. v. i. (ÚŽFG/14/01).

Abstract:

The origin and evolution of functional haemoglobin polymorphism in the bank vole (*Clethrionomys glareolus*)

Bank voles (*Clethrionomys glareolus*) can possess one of two haemoglobin variants differing by a functional amino acid substitution in the beta globin, where 52Ser is substituted by 52Cys. The underlying SNP segregates in both major and minor beta globin genes and the substitution increases the resistance of red blood cells to oxidative damage. It was therefore considered as a possible promoter of the post-glacial replacement of bank voles in Britain where the first colonizing population was replaced by a later arriving one. This example that the current distribution of populations of one species might be a result of selection acting upon genetic differences between them makes the bank vole an ideal model for adaptive phylogeographic studies. I mapped the 52Cys distribution in the bank vole populations across Europe by pyrosequencing on the PyroMark Q24 platform and determined the most likely geographic origin of the allele. Based on data from Sanger sequencing of the entire genes I reconstructed the phylogenetic relationships of the 52Cys and 52Ser alleles from different parts of the distribution area. The results point to possible multiple origins of 52Cys, meaning that the allele is not necessarily a legacy of selection in a single glacial refugium. This is supported by the gene conversion analysis, which suggested that the minor gene can, due to its lower expression and thus lower exposure to selection, serve as a reservoir of 52Cys in a population. The allele can once in a while be transferred by gene conversion from HBB-T2 into the highly expressed HBB-T1 and tested by natural selection for current advantage, and either eliminated from HBB-T1 or fixed. Some populations may thus be more important for a species survival under certain conditions than others, a conclusion changing our perception of the species as a sound conservation unit.

Key words: bank vole, Clethrionomys glareolus, Myodes glareolus, DNA, evolutionary adaptation, GIS, glacial refugia, haemoglobin, protein, ice age, postglacial colonization, gene conversion, pyrosequencing, oxidation stress

Abstrakt:

Původ a evoluce funkčního polymorfismu hemoglobinu norníka rudého (*Clethrionomys glareolus*)

Norník rudý (*Clethrionomys glareolus*) může nést jednu ze dvou geneticky určených variant hemoglobinu, které se od sebe odlišují funkční aminokyselinovou záměnou v beta globinu, kdy 52Ser je nahrazen 52Cys. Zodpovědná nukleotidová záměna (SNP) segreguje jak na majoritním tak na minoritním genu kódujícím beta globin a záměna aminokyselin zvyšuje odolnost červených krvinek vůči oxidativnímu poškození. Je proto považována za možného původce nahrazení jedné kolonizující populace norníka rudého později příchozí druhou populací, ke kterému došlo v Británii během postglaciální kolonizace. Tento příklad toho, že současná distribuce různých populací jednoho druhu může být výsledkem působení selekce na genetické rozdíly mezi jednotlivými populacemi, dělá z norníka ideální modelový druh pro studie adaptivní fylogeografie. Ve své práci jsem pomocí pyrosekvenování na platformě PyroMark Q24 zmapovala rozšíření alely 52Cys v populacích norníka rudého napříč Evropou a určila tak nejpravděpodobnější geografický původ této alely. Na základě dat ze Sangerova sekvenování kompletních sekvencí obou genů jsem zrekonstruovala fylogenetické vztahy alel 52Cys a 52Ser z různých částí areálu. Výsledky poukazují na možný nejednotný mutační původ alel 52Cys, což znamená, že tak nemusejí být pouze dědictvím selekce v jednom glaciálním refugiu. Tuto hypotézu potvrzují výsledky analýzy genové konverze, které naznačují, že minoritní gen díky své výrazně nižší expresi a tím pádem také nižšímu selekčnímu tlaku může sloužit jako jakýsi "zásobník" alely 52Cys v populaci. Alela pak může být čas od času z minoritního genu přenesena do silněji exprimovaného genu majoritního, otestována selekcí na svou momentální výhodnost a buď z HBB-T1 opět vyloučena nebo zafixována. Některé populace tak mohou být důležitější pro přežití druhu pod vlivem určitých podmínek než jiné, což je závěr, který významně mění náš pohled na druh jako smysluplnou konzervační jednotku.

Klíčová slova: norník rudý, Clethrionomys glareolus, Myodes glareolus, DNA, evoluční adaptace, GIS, glaciální refugia, hemoglobin, protein, doba ledová, postglaciální kolonizace, genová konverze, pyrosekvenování, oxidační stres

Contents:

List of abbreviations:

Aminoacids: Ala – Alanine; Cys – Cysteine; Gly – Glycine; His – Histidine; Ser – Serine; Thr – Threonine; Val – Valine

APS – adenosine 5' phosphosulfate

ATP – adenosine triphosphate

BP – years Before the Present

cob – cytochrome b gene of mitochondrial DNA

DNA – deoxyribonucleic acid

dNTP – deoxyribonucleotide triphosphate, where N is: A – Adenine; C – Cytosine; G – Guanine; or T – Thymine

GARD – genetic algorithm recombination detection

GRIP – Greenland Ice Core Project

GSH – glutathione

GSSG – glutathione disulfide

HbF – fast migrating haemoglobin variant

HbS – slowly migrating haemoglobin variant

HbSH – haemoglobin possesing a thiol group

HbSSG – glutathionylated haemoglobin

HKY + I – Hasegawa, Kishino and Yano substitution model with allowance for invariant sites

IPTG – Isopropyl β-D-1-thiogalactopyranoside

K2 + I – Kimura 2-parameter substitution model with allowance for invariant sites

K2 + G – Kimura 2-parameter substitution model with evolutionary rate variation among sites modeled by a discrete Gamma distribution

LGM – last glacial maximum

ML – maximum likelihood

mtDNA – mitochondrial DNA

PCR – polymerase chain reaction

PPi - pyrophosphate

RBC – red blood cells

RLU – relative light units

ROS – reactive oxygen species, the so called free radicals

SBP – single breakpoint recombination

SNP – single nucleotide polymorphism

Taq DNA polymerase – polymerase isolated from bacteria *Thermus aquaticus*

Tm – melting temperature

X-gal – 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

List of figures and tables:

1 Introduction

We are undoubtedly living in a time of changing climatic conditions. This is particularly evident in case of global temperature and hydrological cycle patterns such as air and sea surface temperature, precipitation, evapotranspiration or air humidity (Hartmann et al., 2013). Changes in these environmental variables significantly and visibly influence most species by altering their biology (Walther et al., 2002; Parmesan and Yohe, 2003; Root et al., 2003). Emerging data on adaptive evolution being involved in species response to changing environmental conditions point to significance of intraspecific genetic variability as a substrate for natural selection driven by climate change (Bradshaw et al., 2000; Pulido et al., 2001; Réale et al., 2003; Bearhop et al., 2005; Nussey et al., 2005; Jump et al., 2006). Appreciating such information is of utmost importance for example to draw sound species conservation plans (D'Amen et al., 2013). However, to be able to choose the conservation units correctly or to be able to predict future species fate, it is necessary to understand which parts of the intraspecific variability specifically are involved in species response to climate change. In the species with a short generation interval, such as the common fruit fly (*Drosophila melanogaster*) for example, it is possible to study the impact of the current climate change directly, ideally by a combination of observational and experimental methods with genetics and genomics (Kellermann et al., 2012; Kelly et al., 2012; Lavington et al., 2014; Overgaard et al., 2014; Rane et al., 2015). Nevertheless, for many species including some ecologically important ones this approach is not applicable and then understanding the species reactions to past climatic change is valuable for forecasting the future impacts as well (Stewart et al., 2010). A cornerstone of such research is phylogeography.

Phylogeography, first introduced as a separate field of biogeography in 1987 by Avise and colleagues (Avise et al., 1987), is using molecular markers to gain insights into how and by which processes contemporary geographical distribution of gene lineages was shaped. In other words, in contrast to phylogenetics it is focused on the spatial and temporal dimension of microevolution on intraspecific level (Avise, 1998, 2009). Phylogeographical studies focused on Europe mostly deal with the impact of the Quaternary ice ages and postglacial dispersal of the species into new territories, since

these factors were identified as the most significant for forming the current population genetic structure (Hewitt, 2000).

The importance of the Quaternary period for impacting the within species genetic structuring lies in climatic oscillations characteristic for this era. A dominant feature of the Quaternary period is the repeated dramatic cooling of the global climate paced by eccentricities in the Earth's orbit, resulting in a series of Ice ages with an approximately 100 kyr periodicity (Hewitt, 1999). The Greenland Ice Core Project (GRIP) provides further evidence for climatic oscillations of approximately 7 – 15 °C within the 100 kyr cycle. These changes appeared rapidly within just few decades, lasted for hundreds of years and caused major shifts in species distributions (Hewitt, 2004). The best understood last full glaciation cycle and specifically the Last Glacial Maximum (LGM) lasting from approximately 26500 to 19000 years Before the Present (BP; Clark et al., 2009) is an excellent example of changing climatic conditions impacting species and the global diversity (Barnes et al., 2002). The fact that the origin of genetic variation within most species predates the LGM makes phylogeography a powerful tool for understanding species response to unfavourable climatic conditions during that period.

1.1 Taking selection into account - adaptive phylogeography

Phylogeography certainly helped to understand for example where the European species survived the LGM and how they reoccupied the vacant ecological niches after the onset of warming and retreat of the ice sheet (reviewed by Hewitt, 1999, 2004). Nevertheless, the limits of phylogeography lay in the assumption that the genetic constitution of the current species populations in a particular area simply reflects the genetic characteristic of the first population colonizing that area (Kotlík et al., 2014). In consequence, favourite markers of choice for inferring historical biogeographical scenarios have long been those expected to be selectively neutral, such as mitochondrial DNA (mtDNA). Being small and compact, exhibiting rapid evolution and being easy to isolate and assay, mtDNA meets many of the requirements for an ideal phylogeographical marker proposed by Avise et al. (1987). However, the presumed selection neutrality has currently been questioned (Bazin et al., 2006; Soares et al., 2013), together with other important

mtDNA properties such as purely maternal inheritance or absence of recombination (Kvist et al., 2003; Tatarenkov and Avise, 2007).

The traditional 'selectively neutral' approach in phylogeography is challenged by studies documenting the occurence of a genetic replacement during the postglacial recolonization. This process was most likely driven by the climate change ongoing at that times. Evidence from contemporary populations and ancient DNA shows that a complete population replacement occured and could have been quite common during the postglacial colonization (Barnes et al., 2002; Hofreiter et al., 2007; Searle et al., 2009). When also a replacement during the recent climate change is taken into account (Cook, 1975), it leads to the conclusion that the promoter of such events could have been selection (Kotlík et al., 2014).

If this is true, then the current distribution of populations of a particular genetic constitution may be a result of selection acting on genetic differences among populations of the same species. Thus rather a novel approach of adaptive phylogeography should be adopted. Adaptive phylogeography as defined by Kotlík et al. (2014) aims to shed light on which aspects of the genetic constitution of a particular population may have been favoured during the colonization process. It postulates that the insight into how selection shaped species phylogeographic pattern should be gained not only by collecting data on frequency and distribution of genetic markers, but also by functional analysis of gene products (Kotlík et al., 2014). A strong support for the role of selection is the genetically based physiological divergence between two populations of the bank vole that partly replaced each other during the post-glacial colonization of Great Britain (Kotlík et al. 2014).

1.2 Model species

The bank vole *Clethrionomys glareolus* (Schreber, 1780, also *Myodes glareolus*) is a small rodent species from the family Cricetidae (Fig. 1). Even though it is often confused with the house mouse *Mus musculus* (subfam. Muridae), the bank vole has noticeably more robust body and smaller ears and eyes, which is typical for the Arvicolinae subfamily (voles and lemmings) of Cricetidae. Bank voles differ from most of other vole species by

the significantly reddish-brown fur on their back and whitish underparts. They also have longer tail, which can reach half the length of their body (Macdonald and Barret, 1993).

Fig. 1 Bank vole (*Clethrionomys glareolus*). Photo taken by P. Kotlík.

Inhabiting all kinds of woodlands, typically deciduous or mixed forests but also boreal forests in Scandinavia, bank vole is a typical representative of a temperate forest-dwelling species. It is often present also in parks, scrubs and hedges. It feeds on berries, seeds, roots, nuts, mushrooms, young twigs and also small insects (Macdonald and Barrett, 1993).

What makes the bank vole an ideal model for phylogeographical studies focusing on postglacial colonization of Europe is its distribution range extending to various climatic conditions. It stretches from Great Britain in the west through the whole continental Europe to Altai mountains in Russia in the east and from Mediterranean sea in the south to latitudes beyond the Arctic circle in the north (Fig. 2; the bank vole distribution range is in orange colour).

1.2.1 Key findings derived from mtDNA

In compliance with the general popularity of the cytochrome *b* (*cob*) gene sequence of mtDNA as a vertebrate phylogeographic marker (Johns and Avise, 1998) most of the previous studies using the bank vole as a model based their findings on this marker. Fig. 2. shows nine mtDNA lineages that were identified based on the *cob* data in the studies of Deffontaine et al. (2009, 2005), Kotlík et al. (2006) and Colangelo et al. (2012).

Members of our laboratory (Filipi et al., 2015) revised later this bank vole phylogeography using the complete mitochondrial genome. All the previously described lineages exhibited high support with one exception. The Basque lineage previously identified by Deffontaine et al. (2009) in the border area of northern Spain and southern France was newly defined as a previously undetected nuclear-encoded pseudogene lineage and mtDNA of those bank voles was reassigned to the Spanish lineage (Filipi et al., 2015). Filipi et al. (2015) also expanded the known distribution range of Carpathian clade described by Kotlík et al. (2006) by adding the northern and western periphery of Great Britain and Fennoscandia to it. The Carpathian clade is of utmost importance for elucidating the colonization of Europe after the LGM since its discovery strongly supported the existence of 'northern' glacial refugia (Kotlík et al., 2006).

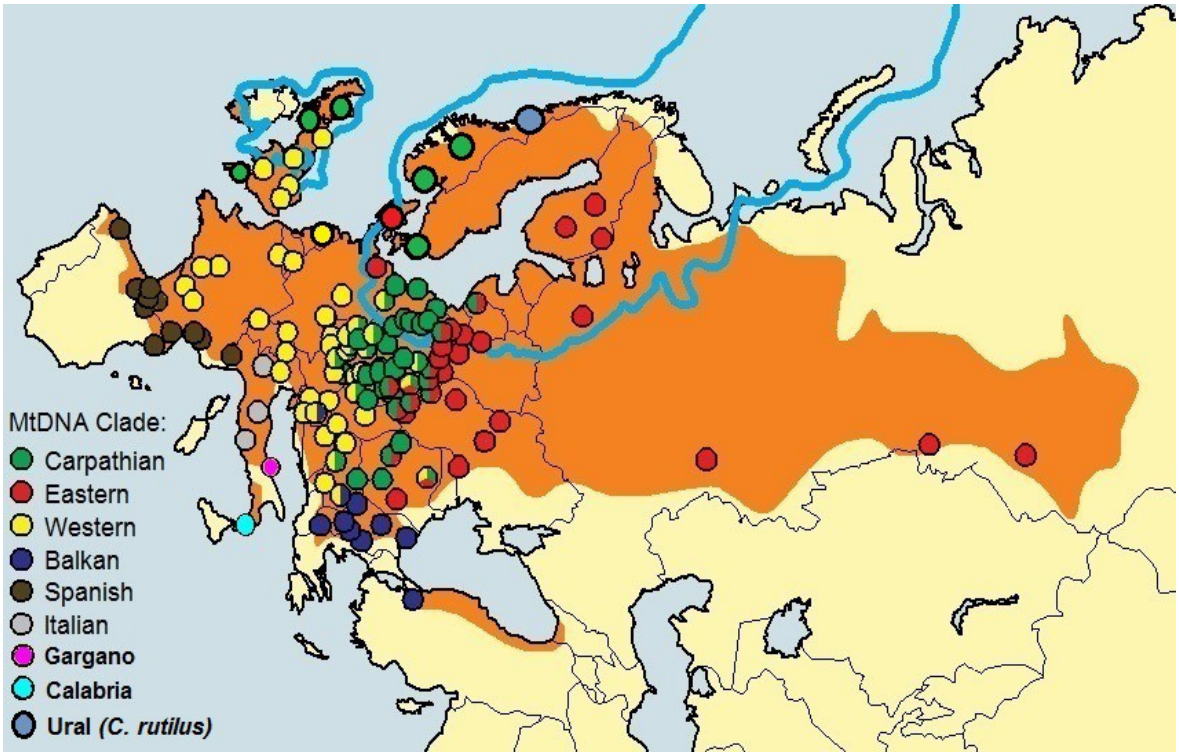


Fig. 2 Bank vole distribution range and identified mtDNA lineages. Blue lines show extent of ice sheets during LGM. Modified from Wójcik et al. (2010) by Karolína Filipi.

1.2.1.1 Evidence for cryptic northern glacial refugia

It is now considered a fact that the LGM had severe impact on species distribution in Europe, especially in higher latitudes. These parts were inhospitable for most of the temperate species due to the presence of ice sheets, permafrost or very harsh climatic

conditions (Hewitt, 2004). A predominant theory has been that the temperate species survived these periods of adverse conditions in areas with milder climate located in Iberia, Italy and Balkans, i.e. the three major Mediterranean peninsulas in southern Europe (Taberlet et al., 1998). The recolonization of Europe after the onset of deglaciation was carried out from these refugia. Patterns of this recolonization were described in a number of species, three of which became a paradigm for others. These were the brown bear (*Ursus arctos*), meadow grasshopper (*Chorthippus parallelus*) and western European hedgehog (*Erinaceus europaeus*) together with its parapatric sister species southern white-breasted hedgehog (*Erinaceus concolor;* for detailed description see Hewitt, 2000). However, phylogeographic studies supported also by fossil evidence came up with findings challenging the universality of this theory.

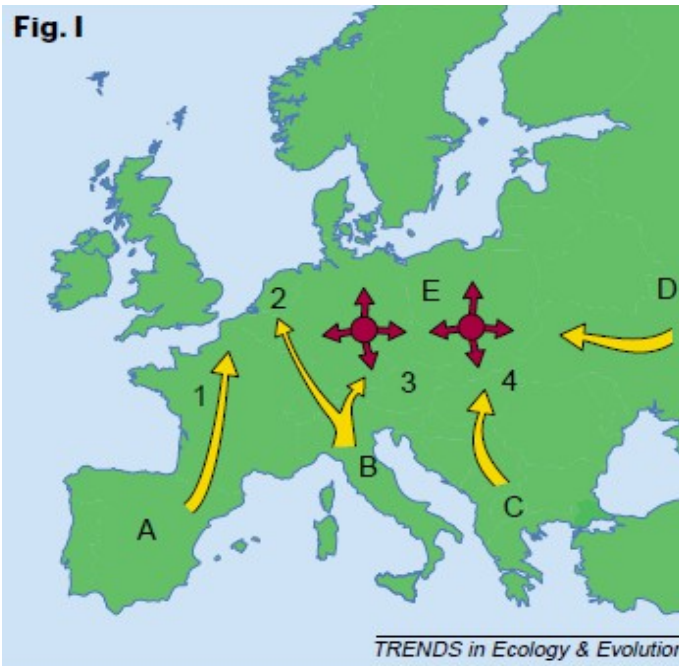


Fig. 3 Location of glacial refugia and possible directions of colonization routes after LGM. 'Traditional' southern refugia are in yellow, alternative cryptic northern refugia in red. Reprinted from Stewart and Lister (2001).

Based on cytochrome *b* sequence data for bank voles and common and pygmy shrews (*Sorex araneus* and *Sorex minutus*, respectively), Bilton et al. (1998) formulated the hypothesis that the Mediterranean populations of these species did not contribute much to the postglacial colonization of Europe. In their review, Stewart and Lister (2001) pointed to possible existence of multiple 'cryptic northern refugia' in central and western

Europe in areas with suitable stable microclimate (Fig. 3). Further analysis of the bank vole molecular data carried out by Deffontaine et al. (2005) pointed to Carpathian mountains as a possible area where the temperate woodland species could have survived the LGM. This hypothesis is supported not only by fossil data (Horáček, 2000) but also by Kotlík et al. (2006) who discovered a new bank vole mtDNA clade in the Carpathian area distinct from the other described mtDNA European lineages. The divergence time and also significant gene flow out of the Carpathians fit perfectly into the scenario of recolonization from these northern refugia (Kotlík et al., 2006). The importance of the Carpathian refugium for the postglacial colonization of Europe (and hence also of the bank vole as a model species) was further emphasized by Wójcik et al. (2010) who discovered the occurence of the Carpathian clade in the area between the Carpathian mountain range and the Baltic Sea and also by the aforementioned findings of Filipi et al. (2015).

1.2.1.2 Celtic fringe – evidence of population replacement

Determining the mtDNA lineages and the discovery of the northern glacial refugium for the bank vole was just a beginning of its fruitful history as a model species. Its excellent properties were used by Searle et al. (2009) who turned their attention to Great Britain and found evidence for a population replacement, findings that challenge the sole use of presumably selectively neutral markers in phylogeography.

The British Isles are a very interesting target for phylogeographical studies. Lowered sea level by approximately 120 m during the LGM (Rohling et al., 1998) resulted in uncovering much more land than today. In some areas, these landmasses served as landbridges. One of such landbridges, Doggerland, was located between the British Isles and Europe (Spinney, 2008). Doggerland enabled species, including humans, to follow the improving climatic conditions and to reoccupy the British Isles previously abandoned because of the ice advance (Barton et al., 2003). However, the warming during the deglaciation lead to rising of the sea level and at approximately 8000 BP (Weninger et al., 2008) the Doggerland was submerged beneath the North Sea. This isolation of Britain from the continental Europe prevented it from being further colonized by small mammals such as the bank vole, leaving the British populations in isolation from any gene flow from the continental populations (Searle et al., 2009).

Searle et al. (2009) carried out a study in which they used three small mammal species as a model to help answer the questions about the recolonization of Britain by humans after the LGM. An interesting feature of human populations inhabiting the British Isles is the culturally, linguistically and genetically distinct Celtic fringe in the western and northern periphery of Britain. The vernacular name 'Celtic fringe' was assigned to the peripheral areas based on traditional but controversial scenario in which Britain was colonized by Celts, who replaced the earlier inhabitants, but later were themselves partially replaced by Anglo-Saxons (Searle and Wilkinson, 1987; Searle et al., 2009). To shed more light on possible processes that lead to the formation of the Celtic fringe, Searle et al. (2009) used mtDNA of the bank vole and two other small mammal species, the field vole (*Microtus agrestis*) and pygmy shrew for a phylogeographic survey. The obtained results combined with data from the water vole (*Arvicola terrestris*) (Piertney et al., 2005) showed a striking pattern. In all studied species, two distinct mtDNA lineages were discovered and their distribution ranges exhibited a clear-cut boundary forming a pattern highly similar to the human Celtic fringe (Fig. 4).

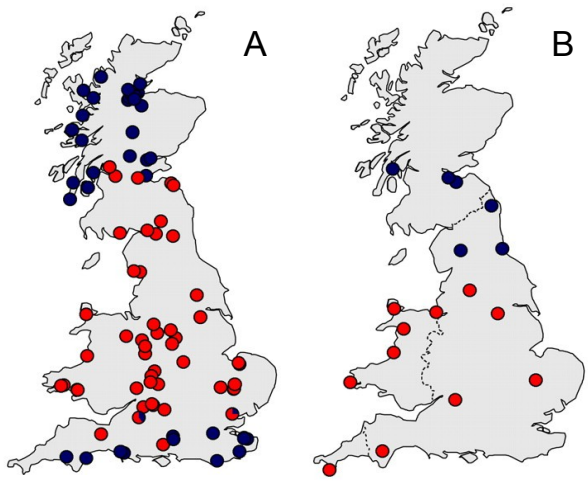


Fig. 4 Celtic fringe pattern of small mammals in Britain represented by **A:** bank vole mtDNA, **B:** bank vole haemoglobin.

Together with the clear north-south trend this pattern lead to the hypothesis of a two phase colonization of Britain over the Doggerland landbridge with a population replacement event. The genotypes currently found in the peripheral areas of western, northern and, in some species, also southern Britain were assumed to be the relict of the first colonizing population which had been partially replaced by another one possibly

coming from a different refugium and currently being distributed in the central and eastern parts of Britain (Searle et al., 2009). These conclusions were later supported by Filipi et al. (2015), who classified the bank vole populations in the periphery of Britain to the Carpathian lineage and the second, central population to the Western lineage.

The origin of the two colonizing populations in two geographically and, what is more important, latitudinally separate refugia points to an idea that the different genetic constitution of these two consecutive populations could have given one of them an adaptive advantage over the other in the changing climatic conditions and that therefore selection could have promoted the replacement (Searle et al., 2009; Filipi et al., 2015). However, no evidence was found that the replacement was promoted by functional divergence of the proteins encoded by mtDNA (Filipi et al., 2015). Thus the source of the assumed adaptive divergence had to be somewhere else in the genome.

Interestingly, the Celtic fringe pattern discovered in the small mammal species by Searle et al. (2009) and Piertney et al. (2005) was also visible in the data from bank vole haemoglobin electrophoresis (see Fig. 4) reported previously by Hall (1979). This surprising concordance of the bank vole haemoglobin with mtDNA markers of the bank vole and other species pointing to the same replacement history attracted the attention of Kotlík et al. (2014). They identified the genetic basis of the two haemoglobin variants and indicated that this protein may have been involved in the population replacement (Kotlík et al., 2014). Since these findings are a cornerstone on which I built my work a separate chapter is dedicated to haemoglobin.

1.3 Haemoglobin coding genes as phylogeographic markers

Haemoglobin, a molecule carrying oxygen from the lungs to the tissues and enabling tissue oxidative metabolism, is a crucial protein component of the vertebrate red blood cells (RBC). Vertebrate haemoglobin molecules are tetrameric, with two α-globin subunits and two β-globin subunits (Fig. 5). These are further composed of 141 and 146 amino acid residues, respectively, and the type and sequence of the residues composing the globin chains largely influences the haemoglobin biochemical properties (Mozzarelli et al., 2013). The primary structure, i.e. aminoacid sequence of haemoglobin is variable in vertebrates on both inter- and intraspecific level. The evidence of haemoglobin

adaptations to deal with extreme environments (i.e. hypoxic waters, high latitudes, extreme temperatures) suggests that the variability of haemoglobin might often be functional and crucial for species survival under certain environmental conditions (see review in Weber and Fago, 2004).

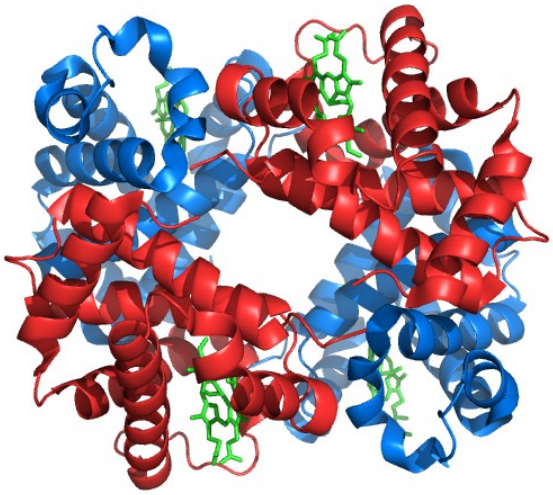


Fig. 5 Model of human haemoglobin based on PDB structure 1GZX, and rendered with PyMol Molecular Graphics System by P. Kotlík. The α-globin subunits are in red, the β-globin subunits In blue. The green parts represent the haem group.

In the particular case of small mammals, the molecular and functional basis of an altitudinal adaptation of haemoglobin has been described for example in the deer mouse (*Peromyscus maniculatus*). The haemoglobin with higher oxygen-binding affinity is favoured in hypoxic areas of higher elevation (Storz et al., 2007). The difference in oxygen-binding affinity is underlaid by five amino acid substitutions in the α-globin and four in the β-globin subunits. Significantly high level of altitudinal differentiation is exhibited by genes coding for the beta globin chains, presumably due to a strong divergent selection. Functional analysis further revealed that there is a difference in the oxygen-binding affinity between the products of the distinct alleles of these β-globin genes (Storz et al., 2009). This evidence highly supports the credibility of the globin genes as a means to study the intraspecific adaptive evolution and the role of selection in shaping the current genetic structure of populations.

1.3.1 Molecular basis of haemoglobin variability in British bank voles

As mentioned before, Hall (1979) reported the existence of two distinct haemoglobin types in Britain with very sharply defined geographical distribution ranges where no locality exhibited presence of both variants at once. The haemoglobin types migrated at different speeds on electrophoresis and this is the reason why the variants were provisionally labeled as HbS (migrating slowly) and HbF (migrating faster; Hall, 1979). This was a rather unlucky choice, since studies dealing fetal haemoglobin refer to it as HbF (for example Dan and Hagiwara, 1967). However, I follow here the Hall's terminology.

There are five globin genes in the bank vole that can be responsible for the variation. Three paralogous genes, referred to as HBA-T1, HBA-T2 and HBA-T3 based on comparison with other rodents (Marková et al., 2014), code the α-globin chains. The two remaining, coding the β-globin chains, have a clear orthology to β-globin genes of other rodent species and are referred to as HBB-T1 and HBB-T2 (Kotlík et al., 2014). However, it is specifically the β-globin subunit, which is significant for the haemoglobin variability as the HbS and HbF variants are distinguished by two distinct β-globin polypeptides as discovered by an electrophoretic analysis (Fig. 6; Kotlík et al., 2014).

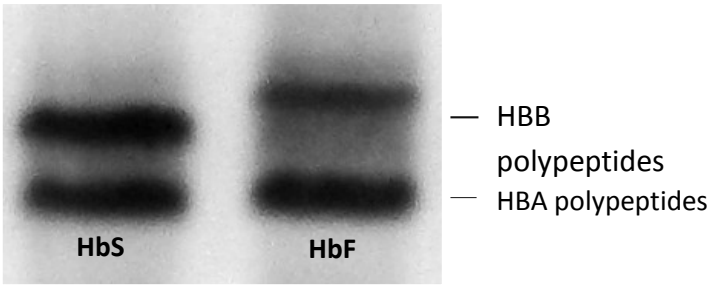


Fig. 6 Urea-cellulose acetate electrophoresis of globin polypeptides demonstrating β-globin differences between HbS and HbF variants. Modified from Kotlík et al. (2014) electronic supplementary material.

The β-globin variability is underlied by the non-synonymous amino acid substitution Ser52Cys, where serine (Ser) is replaced by cysteine (Cys) at position 52 in the protein chain counted in the 5' → 3' direction excluding the start codon. This aminoacid substitution segregates at both the HBB-T1 and HBB-T2 genes and it is a result of a single nucleotide polymorphism (SNP) in the β-globin gene sequence. While the 52Ser is coded by the TCC triplet, the 52Cys is coded by the codon TGC (Fig. 7). In addition to the

Ser52Cys substitution, a second non-synonymous substitution of alanine (Ala) by valine (Val) is located at position 58 (Ala58Val). However, both 58Ala and 58Val are present in HbF as well as in HbS, so this substitution does not relate to the distinction between the two haemoglobin types (Kotlík et al., 2014).

In contrast, the Ser52Cys polymorphism at HBB-T1 exhibits perfect association with the HbS and HbF variants of the British bank voles. The individuals bearing the HbS variant are homozygous for Ser and those bearing HbF for Cys. Animals heterozygous for haemoglobin are, as expected, heterozygous also for Ser/Cys at HBB-T1. Expectedly, the spatial pattern of the 52Ser and 52Cys allele distribution in Britain matches the distribution of HbS and HbF reported by Hall (1979), with the 52Ser allele being present in the north and 52Cys in the south. The only exception is the presence of the 52Ser allele in the southern periphery of Britain in Devon, which was not detected previously by Hall (1979), nevertheless it perfectly reflects the 'Celtic fringe' of the mtDNA distribution reported by Searle et al. (2009) and Kotlík et al. (2014).

In contrast to HBB-T1, no complete association with HbS and HbF variants was detected in case of the same Ser52Cys polymorphism segregating at HBB-T2 (Kotlík et al., 2014). This can be explained by the fact, that HBB-T2 gene exhibits 24-fold lower expression level than HBB-T1, which means that its contribution to β-globin synthesis is only minor and the gene is thus expected to be exposed to less selection pressure than HBB-T1. Thus HbS and HbF variants are result of the allelic variants at site β52 segregating at HBB-T1 (Kotlík et al., 2014). Nevertheless, it is important to note that no HBB-T2 haplotypes are shared between the southern and northern localities, with Devon being an exception again (Kotlík et al., 2014).

Kotlík et al. (2014) hypothesized that the sharing of the Ser52Cys polymorphism by HBB-T1 and HBB-T2 genes is result of gene conversion. Even though they managed to detect gene conversion between the β-globin genes, the conversion tracts did not include one or the other polymorphic amino acid site and thus the mechanism behind the shared polymorphism remains unresolved (Kotlík et al., 2014).

1.3.2 Functional significance

Since the replacement of serine by cysteine is the only difference between HbS and HbF variant it should be logically also the source of any functional difference between the two variants (Kotlík et al., 2014). When it comes to their structure, the only difference between Ser and Cys is that Cys contains sulphur instead of oxygen in Ser (Fig. 7). Such a small change may seem insignificant, however the thiol (-SH) group can become highly reactive under certain conditions and that lends Cys significantly different physical-chemical properties from Ser (Jacob et al., 2005).

Cys thiol group (-SH) ionizes easily forming a deprotonated thiolate anion with a negative charge ($-S^-$). The thiolate anion is able to form disulfide bonds in an oxidation reaction. Intramolecular disulfide bonds are important for protein structure but thiolates are significant for physiology mainly when exposed on surface of the haemoglobin molecule. Intermolecular disulfide bonds can be then formed with other protein and non-protein thiol groups (Reischl et al., 2007; Kotlík et al., 2014).

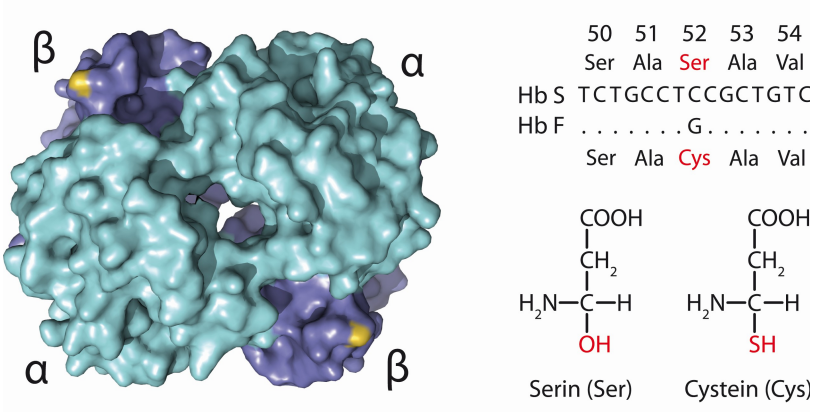


Fig. 7 Model of bank vole haemoglobin showing in yellow the locations of the β52Cys sulphur atoms on its surface (left). Short segment of the β-globin gene showing the SNP distinguishing between Cys/Ser (top right). Structural formulas of Ser and Cys showing their difference (down right). Figure prepared by Petr Kotlík for article in Živa magazine (see Annexe 11).

Cys reactivity depends generally on its position within the haemoglobin molecule and the fraction of the thiolate anion present. The latter is determined by pK_a value which is negatively correlated with the thiolate anion fraction and depends on electrostatic interactions between the anion and its microenvironment (Jacob et al., 2005). For example positively charged residues in vicinity of the thiolate anion can help the Cys to retain lower pK_a values than the intrinsic 9 by stabilizing the anion by hydrogen bond

(Miranda, 2000). In addition to that, the accessibility to solvents affects Cys deprotonation. Thus the exposed position of the Cys on the surface of the haemoglobin molecule is crucial for its reactivity (Jacob et al., 2005).

A possible indication of a reactive Cys present in the haemoglobin molecule is formation of intramolecular disulfide bonds leading to polymerization when exposed to oxidizing agent. In the bank vole the polymerization of HbF was first detected by Hall in 1979 and later verified by Kotlík et al. (2014). In addition to that, Cys in HbF is proved to be completely external (see Fig. 7) with a low pK_a value that appears to be largely determined by a hydrogen bond with the β50Ser side-chain (Kotlík et al., 2014). This means that up to 94% of Cys in red blood cells of bank voles with HbF variant is in the highly reactive thiolate anion form (Kotlík et al. 2014), which is quite unusual in mammals (Reischl et al., 2007).

External thiols are found only in 6 % of mammalian α-globin and in 5 % of β-globin chain sequences. However, only the latter contain 4 % of truly external thiol residues, the rest is present in external crevices as reviewed by Reisch et al. (2007). In their extensive review, these authors studied 155 mammalian species and the totally external thiol occured only in 7 species. These were the rat (*Rattus norvegicus*), guinea pig (*Cavia porcellus*), lesser hedgehog tenrec *(Echinops telfairi*), house shrew (*Suncus murinus*), mouse (*Mus musculus*), virginia opossum (*Didelphis marsupialis virginiana*) and brown greater galago (*Galago crassicaudatus;* Reischl et al., 2007).

Interestingly, one of these species posseses the Cys on the same positon as bank vole, at β52. It is the lesser hedgehog tenrec and the presence of Cys is thought to be possibly related with daily shifts in their activity and associated changes of temperature and metabolic rate (Piccinini et al., 1991). In rat and guinea pig the external thiol groups are hypothesized to be involved in antioxidant defense of red blood cells and detoxication of the organism from carcinogenic electrophiles (Rossi et al., 1998; Miranda, 2000).

1.3.3 Antioxidative effect of cysteine in haemoglobin

Red blood cells functionate as oxygen carriers in the organism and as such are under constant risk of oxidative damage due to reactive oxygen species (ROS) formation. To mitigate the negative effect of ROS enzymatic and non-enzymatic antioxidative

defenses are of utmost importance. Among the non-enzymatic antioxidants one of the most important and abundant is glutathione (GSH), a tripeptide formed by three amino acids including one cysteine (Reischl et al., 2007). When the level of ROS production is increased, glutathione is capable of reducing the ROS while being oxidized itself producing glutathione disulfide (GSSG; Filipovska and Murphy, 2006). The reduced form (GSH) is more abundant than the glutathione disulfide in the intracellular environment but the GSH/GSSG ratio is lowered under oxidative challenge. Since the reducing potential of the red blood cells is dependent on the GSH/GSSG ratio, the regeneration of the glutathione pool is crucial for effective antioxidant protection (Reischl et al., 2007). The GSSG can be reduced back to GSH by enzymes, however another pathways are also possible. Haemoglobin possesing a thiol group (HbSH) can help regenerate the GSH pool by reacting with glutathione disulfide, forming glutathionylated haemoglobin (HbSSG) and releasing one molecule of GSH as described in rats (Rossi et al., 1998). In addition to that, the data on rat haemoglobin show that the thiol group can react with ROS directly and that the haemoglobin possessing highly reactive Cys can intercept the ROS even more efficiently that GSH (Rossi et al., 1998).

The ability of bank vole haemoglobin HbF to intercept the ROS was tested experimentally by Kotlík et al. (2014) by a luminometric method to measure the 'Total Radical-trapping Antioxidant Potential', so called TRAP test.

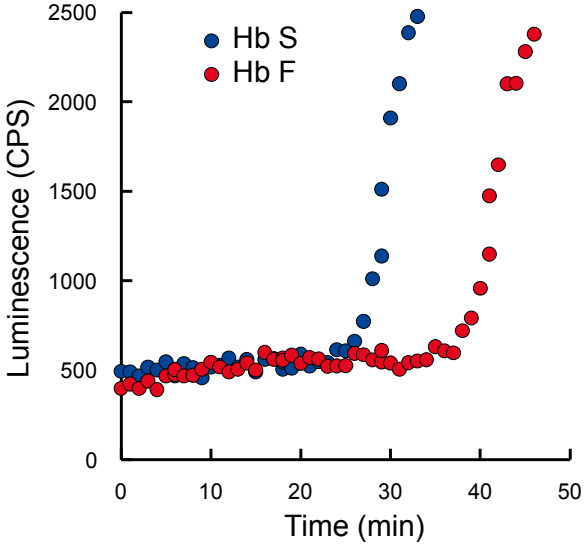


Fig. 8 Antioxidant capacity of RBCs of bank voles carrying HbS and HbF represented by results of the TRAP test. HbF significantly delays the luminol oxidation.

TRAP test uses organic radical producer and measures the delay in oxidation of luminol to determine the period during which the red blood cells are able to capture ROS. Luminol, when oxidized by ROS, emits a luminescence signal measured by luminometer and thus sudden increase of luminescence means that the antioxidant capacity is depleted. The longer the lag phase until the onset of luminescence, the higher the antioxidant capacity of the sample (Kotlík et al., 2014). Fig. 8 shows clearly, that the samples of HbF haemoglobin with 52Cys exhibited much higher (three fold when related to a standard, trolox) antioxidant capacity than HbS samples. This means that the reactive thiol group present in the bank vole haemoglobin HbF increases the resistance of their red blood cells to oxidative damage (Kotlík et al., 2014).

1.3.4 Evolutionary origin and significance

By comparison with continental populations founded by the same two colonists as the British populations (from Sweden and Netherlands, respectively), Kotlík et al. (2014) demonstrated that the first colonists of Britain carried the 52Ser haemoglobin variant and the second colonists the 52Cys haemoglobin. Filipi et al. (2015) then provided more comprehensive data on mtDNA variation linking the first colonists to the Carpathian clade and the second colonists to the Western clade.

These findings again support the idea of two phase colonization of Great Britain as proposed by Searle et al. (2009). However, the two successive populations did not differ only in their mtDNA, but also in their haemoglobin type (Kotlík et al., 2014). The higher resistance to oxidative stress might be advantageous under conditions with higher ROS production, such as higher muscular activity, increased growth rate, increased reproductive effort or thermal stress. This means that the physiological difference between the two haemoglobin variants might lead to different performance of the populations under various conditions. In the case of the British bank voles the different performance during divergent stages of the postglacial climate change could be the reason for the population replacement (Kotlík et al., 2014). It is possible, that the first colonizing population coming from the northern refugium in the Carpathians was better adapted for colder climate that prevailed during the early stages of the deglaciation, but that it was outperformed by the second population coming from a more southern

refugium as the climate became warmer. Taking the HbF properties into account, the advantage could have been for example a possible higher reproduction rate of individuals carrying HbF afforded by the increased antioxidative capacity of their red blood cells (Kotlík et al., 2014). This means that even though we are not able to establish a direct link between climate and the haemoglobin variants, the possibility that one variant could have been advantageous under certain climatic conditions than the other points to an idea that some populations could be more important for the species survival certain climatic conditions than other populations, depending on their specific genetic constitution.

In my study, I decided to go beyond the boundaries of Great Britain and to map the geographic distribution of β52Cys allele in Europe in order to determine its distribution on the continent and its possible geographic origin and to reconstruct its evolutionary history. Since the amino acid substitution is present in both major and minor β-globin genes, but in the latter is rather functionally insignificant due to the gene's low expression level, I decided to try also to shed more light on the possible influence of gene conversion on such arrangement. Specifically, I wanted to find out if there is evidence of gene conversion that could transfer the adaptive β52Cys allele between HBB-T1 and HBB-T2, which would change its exposure to selection pressure (presumed to be stronger in high expressed HBB-T1 than in low expressed HBB-T2). As a means to obtain the required data, different sequence determination techniques were chosen.

1.4 Sequencing methods

Different questions require different data and approaches to answer them. While to determine the geographical distribution of the Cys allele only the information on genotype on the target SNP position from multiple individuals is sufficient, phylogenetic tree reconstruction requires information on longer gene regions. In case of this study, it would be whole sequences of the major HBB-T1 and minor HBB-T2 beta globin genes, including their three exons and two introns.

Owing to technological development in last 50 years, there are many techniques of DNA sequence determination available. These technologies differ in suitability of their application due to different read length, accuracy and also cost and laboriousness (Ahmadian et al., 2006). Sanger sequencing, a once-revolutionary approach based on

dideoxy chain termination, which was first introduced in 1977 (Sanger et al., 1977), is still very popular. Having gone through significant improvements during the years, it became a robust technique offering rather long reads (500 – 1000 bp) and the ability of reliable de novo sequencing (determination of unknown sequence) (Ahmadian et al., 2006). I thus decided to rely on Sanger sequencing to collect data on the whole beta globin gene sequences. However, the laboriousness of the method and longer time needed for the analysis makes it rather ineffective in case of a single SNP genotyping (Ahmadian et al., 2000). Therefore pyrosequencing performed on Qiagen PyroMark platform (Valencia, CA, USA) was selected for genotyping the variable position of interest. Since the method represents a core element of my research, but is still somewhat less well known than Sanger sequencing, I describe it in more detail in following section.

1.4.1 Highlights and drawbacks of pyrosequencing

Pyrosequencing is a technology that can be used in both confirmatory and de novo sequencing (Ronaghi, 2001). Its limited read lenght makes it ideal for applications where where only a short sequence needs to be determined, such as SNP genotyping (Keller et al., 2011), resequencing of diseased genes (Sha et al., 2012) and microbial typing (Sundquist et al., 2007). Rapid sample preparation, possible high throughput, programmable nucleotide dispensation order and real-time result monitoring are amongst the greatest advantages of pyrosequencing (Fakruddin and Chowdhury, 2012).

The main advantage for SNP genotyping specifically is that the pyrogram readout enables clear distinction of various genotypes as each allele combination results in a specific pattern (Ahmadian et al., 2006) as demonstrated in Fig. 9.

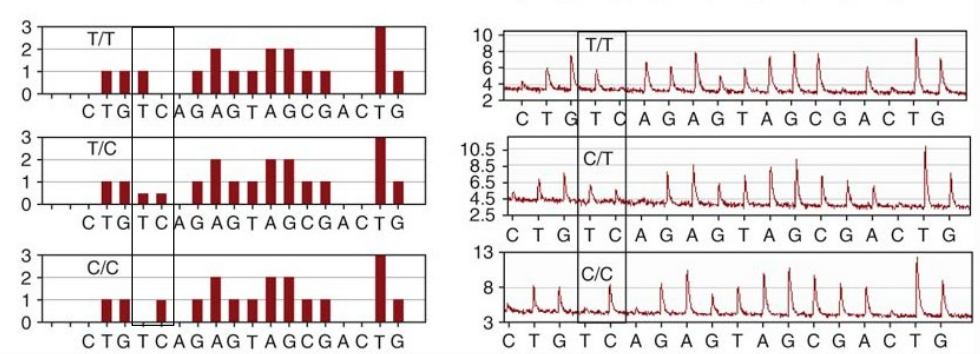


Fig. 9 Expected and obtained pyrograms for T/C SNP genotyping in heterozygous and homozygous individuals. Modified from Royo et al. (2007).

Moreover, pyrosequencing provides not only the data on genotype, but also quantifies the proportion of variants present in tested sample (QIAGEN, 2010). In case of an SNP with two alleles, these values can be used as an indicator of accuracy of the genotype call. In addition to that, software provided with the platform is able to detect various problematic conditions like wide or low peaks, high background, unreliable surrounding sequence pattern and others. This helps to find and eliminate erroneous results.

The PyroMark Q24 platform has the capacity to perform 24 separate assays in one run. Each sample is handled as a separate assay, which means that 24 different samples can be analysed in one run (QIAGEN, 2010).

1.4.2 Pyrosequencing technology and workflow

Pyrosequencing is based on the sequencing-by-synthesis principle and relies on detection of pyrophosphate released during the DNA synthesis (Ronaghi, 2001). Here I describe the application of the pyrosequencing method as used for SNP genotyping with the Qiagen PyroMark platform. My description is based on the information of the Qiagen company published at their website (QIAGEN, 2010).

First, the DNA sequence to be analyzed is amplified by polymerase chain reaction (PCR). A biotinylated amplification primer is used for the strand meant to be the pyrosequencing template (Fig. 10).

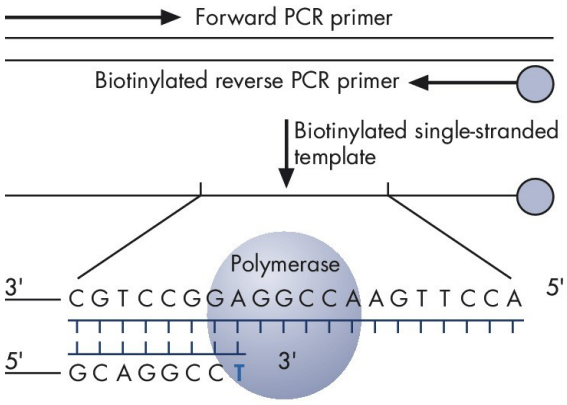


Fig. 10 Schematic representation of amplification of the DNA template and sequencing primer annealing. Copied from (QIAGEN, 2010)

The PCR products are mixed with binding buffer and streptavidin-coated sepharose beads. Since biotin binds to streptavidin, the biotinylated amplicons bind to the beads. The amplicons bound to the beads are then immobilized on filter probes of a vacuum tool. Vacuum applied to the tool creates a current through the probes when introduced into a liquid. Sepharose beads are captured on the filter, since they are too large to pass through. The vacuum tool is then inserted into a series of troughs containing denaturation solution and washing buffer. After denaturation, the DNA strands separate and the one lacking biotin is washed away by the washing buffer. The result is a single-stranded purified DNA template bound to the streptavidin-coated bead.

The single-stranded PCR amplicons are released from the beads and hybridized with the sequencing primer (Fig. 10). Nucleotides (dNTPs, one at a time), enzyme solution and substrate solution are loaded into correct department of PyroMark Q24 Cartridge. Enzyme solution contains DNA polymerase, ATP sulfurylase, luciferase and apyrase. Substrate solution is comprised of adenosine 5' phosphosulfate (APS) and luciferin. During the reaction, all these components are dispensed from the cartridge into the reaction plate with the solutions containing the different nucleotides (A, T, C, G) being dispensed according to a dispensation order defined during the assay setup. When a nucleotide is incorporated in the sequence light signal is emitted and detected.

In case the nucleotide is complementary to the base in the template strand, DNA polymerase catalyzes nucleotide addition to the sequencing primer. Whenever the incorporation happens, pyrophosphate (PPi) is released (Fig. 11). The quantity of PPi released is equimolar to the amount of incorporated identical nucleotides.

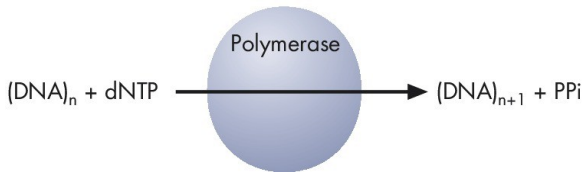


Fig. 11 Schematic image of the pyrophosphate release.
Copied from (QIAGEN, 2010).

PPi is converted to ATP which drives the conversion of luciferin to oxyluciferin. This generates visible light in amounts that are proportional to the amount of ATP (fig. 12). The

light produced is detected by camera device and seen as a peak in Pyrogram. The height of each peak is proportional to amount of the incorporated nucleotides.

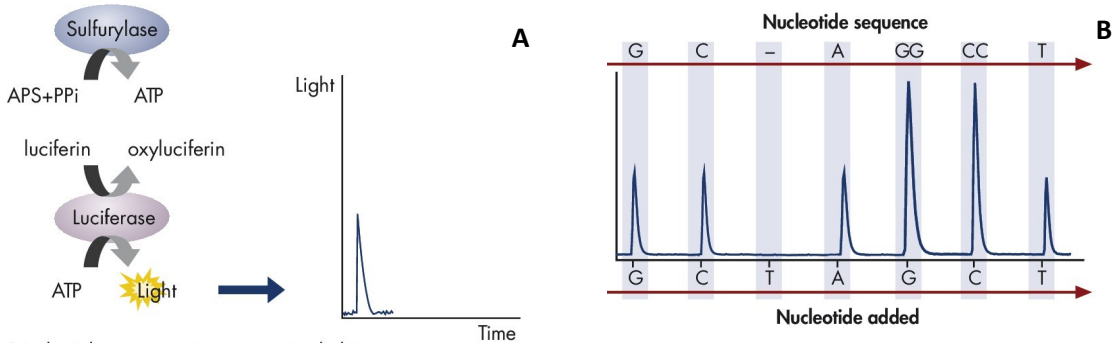


Fig. 12 A Schematic representation of the enzymatic cascade resulting in light emission and peak detection. **B** Schematic representation of the resulting pyrogram. Copied from QIAGEN (2013).

Unincorporated nucleotides and ATP are degraded by apyrase. Another nucleotide is not added to reaction untill the degradation is complete.

2 Aims of thesis

The functional difference between the two bank vole haemoglobin variants is caused by a single non-synonymous mutation in the major β-globin gene. My thesis aimed to genotype the underlying SNP, describe and analyze its distribution in bank vole populations across Europe and use this information to determine the geographic origin of the mutation and its evolutionary history. To meet the goals, pyrosequencing and Sanger sequencing were used in combination with population genetics, spatial analysis techniques in GIS and phylogenetics.

The specific objectives were:

- to develop and optimize pyrosequencing assays to genotype the target variable position separately in the major HBB-T1 gene and the minor HBB-T2 gene
- to genotype the SNP in 500+ specimens from various European populations
- to calculate, analyze and visualize the population allele frequency to determine the possible geographic origin of the mutation
- to sequence representative alleles from different populations using Sanger sequencing
- to clone PCR amplicons for individuals with unphased genotypes
- to construct phylogenetic trees from the resulting sequences to unveil the evolutionary history of the mutation and to estimate whether the 52Cys alleles are identical by descent (have a common ancestor)
- to test for gene conversion between the HBB-T1 and HBB-T2 genes as a possible means of transfer of 52Cys between them

3 Materials and methods

3.1 Sample selection

A total of 550 samples were selected from an extensive inventory of available specimens in our laboratory stock. Both DNA isolates and ethanol preserved tissue samples were taken into account.

The samples were chosen according to the location of the sampling site in such a way that the whole species distribution range was covered as evenly as possible, including the putative glacial refugia (Kotlík et al., 2006). Representatives of all the described bank vole mtDNA clades (Wójcik et al., 2010; Filipi et al., 2015) were included. To be able to estimate the population allele frequency, sampling sites with sufficiently high number of samples were selected preferentially, aiming at 10 voles per locality, and further handled as separate regional populations. Where not enough individuals were available from the same site, specimens from different but geographically close sampling points were included into one 'regional' population. Regional populations are depicted in Fig. 13, for the list of localities, their coordinates and number of samples included see Annexe 1.

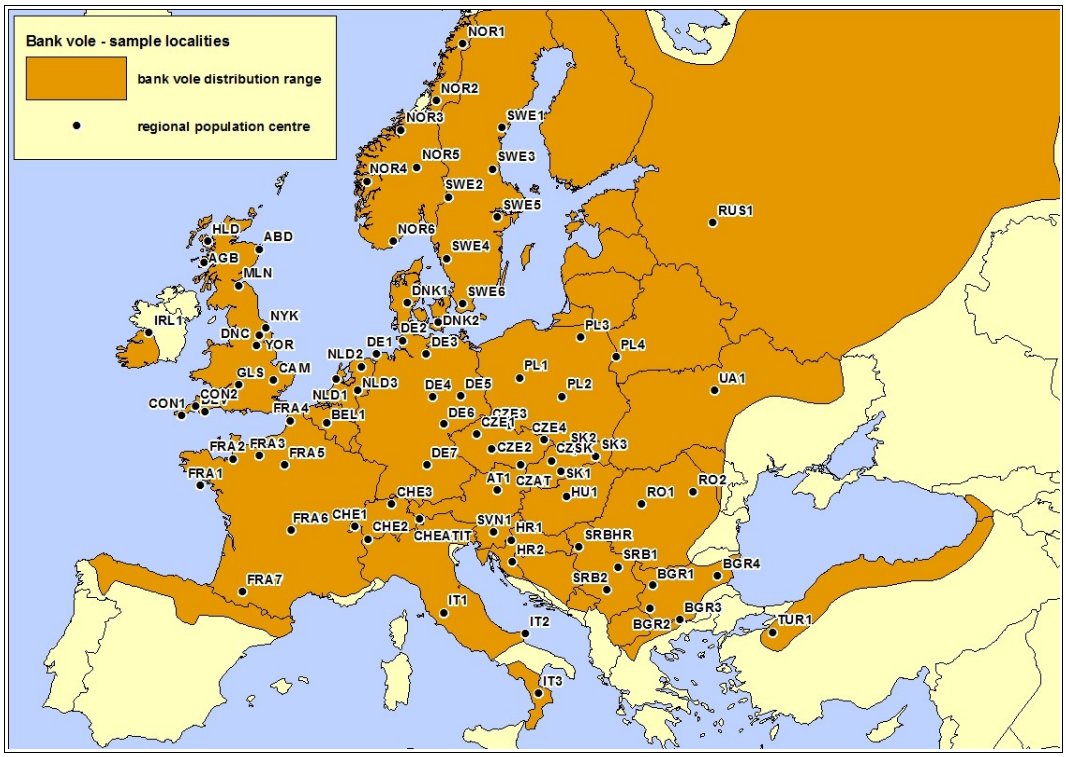


Fig. 13 Locations of the bank vole regional populations. Dots represent the centres of population samples. In case of populations including individuals from two or more sampling sites, location of the centre was calculated as weighted average of sampling sites locations. Bank vole distribution range taken from Wójcik et al. (2010).

3.2 DNA isolation

Total genomic DNA was isolated from 103 ethanol preserved samples using the Qiagen (Valencia, CA, USA) DNeasy Blood and Tissue Kit. When available, liver or spleen tissue was used to ensure optimal yield and quality of the isolated DNA. However, toe, tail or ear clips were used in some cases. Optimized protocol provided with the kit was followed with two specific modifications. To ensure complete lysis, samples were lysed overnight at 55°C, rather than for the recommended 1 to 8 hours (depending on the tissue type). To increase the final DNA concentration in the eluate, only 100 µl instead of 200 µl of Buffer AE was used in the elution step.

Centrifugation steps of all procedures requiring it, including the DNA isolation and further mentioned DNA purification, were performed on centrifuge Mikro 200 (Hettich Zentrifugen).

3.3 Gel electrophoresis

Presence of DNA in the eluate and the degree of its degradation was verified by gel electrophoresis. An agarose gel was prepared by dissolving agarose (SERVA, *research grade* quality) in 0.5% TBE buffer and short boiling in a microwave. The final weight/volume percentage concentration of the agarose was 0.8 %. GoldView™ (Viswagen Biotech) dye was added before pouring the boiled agarose into the gel form. The gel was then left in the form to solidify for 20 min at room temperature.

An aliquot of 5 µl of the DNA eluate was dyed using 1 µl DNA Gel Loading Dye (Thermo Fischer Scientific) and loaded to the gel. Apart from dying the DNA to visualize the result, the loading dye serves also to put weight to the sample to ensure it will sink to the bottom of the well. The gel electrophoresis was run on Labnet Power Station 300 in TBE buffer at room temperature. The voltage for first 20 min was 80V and 98V for next 40 min.

Results of the gel electrophoresis were checked and photographed on transluminator MultImage™ Light Cabinet (Alpha Innotech Corporation).

3.4 Pyrosequencing

For genotyping of the samples, PyroMark Instrument and software was used (Qiagen, Valencia, CA, USA). This method is not only able to determine genotype for large amount of individuals in short time, but also enables to quantify the accuracy of genotype calls.

3.4.1 Pyrosequencing assay design

In a first step a pyrosequencing assay had to be developed. Because the aminoacids Cys/Ser segregate at the same site in both genes coding for beta globin chain in the bank vole (Kotlík et al., 2014) two separate assays were designed - one for the major HBB-T1 gene and second for the minor HBB-T2 gene.

Each pyrosequencing assay is comprised of two amplification primers used for PCR and one sequencing primer. To make the assays gene specific, amplification primers had to be designed in regions where the paralogous genes are most dissimilar. Since the genes differ most in the 3' untranslated region (UTR), the reverse amplification primers were designed in that area. As the recommended length of the PCR amplification product used for pyrosequencing reaction is only up to 500bp, the forward amplification primers were located as close to the target variable position as possible to produce the shortest possible amplicon. Nevertheless, because the target SNP is located in the 2nd exon on the position 264 and the 3'UTR starts behind the position 1100, the final PCR product could not be shorter than 900bp.

To generate sets of compatible amplification and sequencing primers according the previously mentioned criteria, PyroMark Q24 Assay Design Software was used. The program scores the designed sets of primers according to several parameters specific for pyrosequencing analysis. In that way the quality of pyrosequencing results that can be achieved by each set is quantified. All generated possible primer combinations were manually checked, sets with the forward aplification primer positioned in an intron were discarded. Introns, as non-coding regions, accumulate more variability than exons. Thus placing the entire amplification primer in this region could cause a decrease of the amplification efficiency when amplifying variable sequences. In our samples a rather high

variability in the introns was expected, since the specimens originated from various glacial refugia.

Three primer sets with the best score were then selected for each gene with preference for sets with a forward sequencing primers. If using reverse sequencing primers, the second known SNP on position 282 causing Ala/Val substitution in HBB-T1 gene (Kotlík et al. 2014) would not be scored. This would be due to the location of the sequencing primer, which has to be within 5 bp from the target SNP on position 264 to ensure high accuracy of the genotype call.

The selected sets of primers were evaluated with OLIGO Primer Analysis Software. To avoid non-specific amplification or poor priming efficiency, optimal primer set should generally contain primers with lowest possible mispriming efficiency, high internal stability at the 5' end and low at the 3' end (to avoid extension from a primer annealed only at its 3' end), and most similar melting temperatures of the amplification primers (Rychlik, 2007). The main advantage of OLIGO over PyroMark software is its ability to provide more detailed analyses with quantification of those criteria. Since some of the sequencing primers generated by PyroMark software did not exhibit optimal internal stability profile, OLIGO was used also to design possible alternatives.

3.4.2 Testing and optimization

The assays developed by PyroMark as well as the additional primers generated by OLIGO were tested for functionality and efficiency using samples of known genotype (determined by Sanger sequencing; Kotlík et al., 2014) Representatives of all three possible genotypes were chosen to test power of PyroMark Q24 to distinguish clearly between samples homozygous for Cys, homozygous for Ser and heterozygous.

3.4.2.1 Testing of amplification primers - PCR

First, the amplification primers had to be tested. To test the functionality of each set, the first PCR was always run at the temperature suggested by PyroMark Assay Design software as the optimal Tm. A second PCR was then run with a temperature gradient to find the optimal annealing temperature for each set. Temperatures slightly exceeding the minimal and maximal Tm recommended by OLIGO were taken as the limit temperatures for each gradient. The intermediate temperature steps were calculated automatically by

the C1000™ Thermal Cycler from Bio-Rad Laboratories (Hercules, CA, USA), which was used for all PCR amplification mentioned in chapter *3. Materials and methods*. The highest temperature at which the desired fragments were amplified in sufficient quantity was selected for each primer set as the annealing temperature for further optimization steps.

For the PCR mix a pre-mixed PPP Master Mix (Top-Bio, Prague, Czech Rep.) was used. It is an ideal option for routine PCR analyses since it contains nucleotides, magnesium, Taq DNA polymerase and optimalised reaction buffer in double concentration. For the reaction preparation only the template DNA, primers and PCR H_2O have to be added. The Master mix also contains a loading dye and additives that make possible immediate loading of the samples to gel for electrophoresis. For example of PCR mix and programme see optimized HBB-T1 and HBB-T2 protocols in Annexe 2.

Purity and yield of all performed PCR was verified by gel electrophoresis. An aliquot of 1 µl of the PCR product was loaded to gel. No additional loading dye was added to the sample since PPP Master mix was used for PCR. To be able to verify the size of the amplified fragment, 3 µl of a size marker GeneRuler 100 bp Plus DNA Ladder (Thermo Scientific) was loaded to the gel together with the samples. Other procedures and conditions were same as described in chapter *3.3 Gel electrophoresis*.

3.4.2.2 Testing of sequencing primers

For the first testing purification and pyrosequencing run all procedures and conditions were followed as recommended by Qiagen in PyroMark Q24 Vacuum workstation quick start guide (Annexe 3). The results obtained under these conditions were correct, but marked as low quality due to low peak height (Fig. 14).

The PyroMark software evaluates the results and marks them as PASS, CHECK or FAIL quality primarily according to the peak height which is expressed on the X axis of each pyrogram in relative light units (RLU). CHECK quality requires single peak height at least 10 RLU and PASS at least 20 RLU. These threshold values can be adjusted, however, too low peaks can lead to erroneous results. Other criteria that can affect the quality of the read is signal to noise ratio, width of the peaks and surrounding sequence pattern.

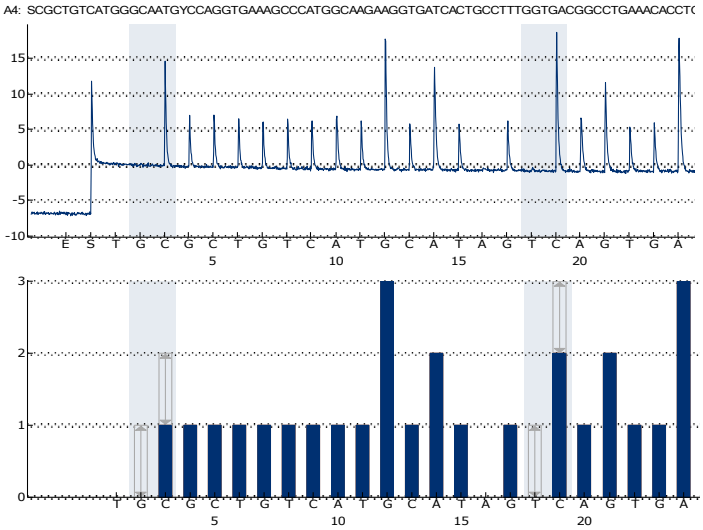


Fig. 14 Pyrogram obtained respecting recommended conditions (top) and expected pattern (down). The Y axis in the top is in RLU, the Y axis in the bottom panel shows number of incorporated nucleotides. Variable positions are in light blue. Empty light grey rectangles show variants of the variable positions. The obtained pyrogram matches the expected pattern and the result is thus correct despite the low peak height.

3.4.2.3 Optimization

An optimization was done to reach adequate quality of the genotypes. The optimization process included testing various procedure conditions and manipulating the amount of chemicals used.

The first assumption was that the amount of DNA in samples used for the pyrosequencing reaction has been too low. This hypothesis was tested by performing a run with DNA concentration gradient from 5 to 30 ng/µl. The DNA concentration was estimated using the GeneRuler 100bp Plus quantification ladder on gel electrophoresis after PCR amplification. When needed, the DNA was diluted to the required concentration by high purity water used for PCR mixes. Other conditions of sample preparation and analysis were left unchanged. Peak height showed tendency to improve slightly with higher concentration but the result was still marked as low quality due to low peak height. Then the DNA concentration was increased up to 50 ng/µl but without any significant improvement of the peak height.

As a second, hypothesis was tested that vortexing PCR products at 1400 rpm during the immobilization step leads to breaking of the long PCR amplicons and thus to low quality results. Lower rpm (1000) for longer time (20 min) was used, with other

conditions unchanged. However, the results showed that lower rpm lead to poor mixing of the PCR producst with streptavidin-coated Sepharose beads. The peak height was even lower than in the first testing run.

Poor results when the mixture was agitated more gently lead us to a hypothesis, that the volume of Streptavidin Sepharose beads added to our PCR product has been too low. The long PCR amplicons could have been sort of a barrier for the beads, preventing the beads, if present in too low number, from mixing properly with the PCR product. To test this hypothesis, a run with samples mixed with 2 µl of Streptavidin Sepharose beads at 1400 rpm for 20 min was performed. The peak height almost doubled and all results were qualified as PASS quality.

Performance of the tested assays was compared. The assay with best and most balanced peak height in all three tested genotypes was selected for each gene.

These final pyrosequencing assays were once more tested in a run where 3 µl of Streptavidin Sepharose beads were added to the PCR products. The increased amount resulted in slightly higher peaks than in the run with 2 µl of beads and was thus incorporated into the final protocol (see Annexe 4a).

3.4.3 Genotyping the samples

The sample processing was divided into 2 consecutive days. One day the selected gene was amplified using the optimized PCR conditions and selected primer set (Annexe 2) and a gel electrophoresis was run to verify the product quality. The second day the PCR products of sufficient quality were processed according to the optimized immobilization, sequencing primer annealing and pyrosequencing run preparation protocols (see Annexe 4). This helped to maximize the efficiency of the pyrosequencing by minimizing the chance for DNA degradation of the PCR products. When possible, 120 samples, with some extra to compensate for potentially low quality ones, were amplified by PCR in one day. This was because the amount of the PyroMark Q24 Gold Reagents is designed for five pyrosequencing runs of 24 samples each. To avoid possible degradation of the reagents by repeated thawing and freezing, the best approach was to prepare and perform all five runs in one day. Subsequent data analysis of the obtained results is described in chapter *3.7.1 Pyrosequencing data analysis*.

3.5 Sanger sequencing

For the phylogenetic tree construction it was necessary to obtain sequence data for the whole beta globin genes. The goal was to compare the phylogenetic relationships of β52Cys and β52Ser alleles. Specifically whether the Cys alleles from different parts of the distribution area are evolutionarily closer to each other than to Ser alleles from the same or other populations. For this purpose representatives bearing the β52Cys allele from different parts of its distribution area were selected, including the isolated areas of higher β52Cys allele frequency in Sweden and the Balkans. To compare the phylogenetic relationships, representatives of the β52Ser allele were added to the list. In total 38 representative specimens were selected for each gene based on the results of allele frequency interpolation, however sequences for some of them were already available from previous studies. Thus, the sequencing analysis was done in total for 30 samples for the HBB-T1 gene and 28 samples for the HBB-T2 gene For list of samples see Annexe 5.

3.5.1 PCR amplification

The major HBB-T1 gene was amplified from the selected samples using PCR primers BT1F1 and BT1R593. For amplification of the minor HBB-T2 gene primers HBB10U19 and HBB1266L21 were used. Both of these sets were originally designed by Kotlík et al. (2014) for the purpose of a previous study in such a way that each of the sets selectively amplified only one of the two genes. For primer sequences, PCR mixes and programmes used see Annexe 6a.

3.5.2 Purification of PCR products and quantification

The obtained amplicons were purified using a commercial kit (QIAquick PCR Purification Kit, Qiagen) and quantified on gel electrophoresis. An aliquot of 1 µl of PCR product was mixed with 0.5 µl of DNA Gel Loading Dye and loaded to gel. In addition, 3 µl of a size marker GeneRuler 100 bp Plus DNA Ladder were loaded to the gel together with the samples. Other procedures and conditions were the same as described in chapter *3.3 Gel electrophoresis*.

3.5.3 Sequencing

All successfully amplified and purified amplicons were sequenced using the Sanger sequencing method. Sequencing primers used for HBB-T1 were the same as for the amplification, for HBB-T2 primer HBB597U19 was used instead of HBB1266L21 (for the primer sequences see Annexe 6b). Both forward and reverse sequencing primers were used since the amplicon length was too high for one read only. Also the accuracy of the reads is lowering with increasing distance from the priming site. All sequencing was done at Macrogen Inc. (Soul, South Korea). For subsequent data analysis see *3.7.2 Sequencing data analysis*.

3.6 Cloning

The haplotype phase for individuals with more heterozygous positions than one was primarily determined by computational methods (see below). Where the phase was not determined with sufficient probability or in sequences with indels present an experimental approach of phase determination, cloning, was adopted. In total 19 sequences were used for cloning. Gene sequences from these samples were amplified again using the same PCR primers, mixes and programmes as for Sanger sequencing. Obtained amplicons were purified using the commercial QIAquick PCR Purification Kit and cloned using Qiagen PCR Cloning kit.

3.6.1 Cloning procedure

Following the procedures as given by Qiagen in PCR Cloning kit handbook the amplicons were ligated into the pDrive cloning vector. Ligation phase was extended to 2 hours as recommended to ensure higher number of recombinants. The pDrive vectors were transformed into competent cells *E.coli lacZΔM15* by heat shock. Cells were plated on media containing antibiotics, specifically ampicilin, enabling selection on cells containing plasmid. IPTG and X-gal solution was spread over the media before plating to enable selection on cells containing recombinant plasmid. Colonies of cells with no plasmid will not grow on ampicilin, colonies of cells containing plasmid without incorporated insert turn blue during cultivation.

3.6.2 Plasmid DNA isolation and verification

Plasmid DNA was isolated from colonies using the commercial QIA Prep® Miniprep Kit from Qiagen. Verification of inserted fragment length was done by restriction of the plasmid by endonucelase EcoRI. This enzyme cuts the plasmid at sites to the left and to the right of the insert so the verification was done by analysis of the DNA fragments by gel electrophoresis.

An aliquot of 1 µl of the digested plasmid DNA was mixed with 0.5 µl of loading dye and loaded to gel. For size verification 1 µl of dyed purified PCR product of cloned samples and 3 µl of size marker GeneRuler 100 bp Plus was loaded to gel too. Other procedures and conditions were the same as described in *3.3 Gel electrophoresis*.

3.6.3 Sequencing

Plasmid DNA of clones containing fragment of the correct length was sequenced using primers M13F20 (5'GTAAAACGACGGCCAGT3') and M13R (5'AACAGCTATGACCATG3'). These primers anneal to the plasmid DNA and span the inserted fragment. Six clones per each amplicon were sequenced when enough recombinant clones were available. Again, for Sanger sequencing the service of the Macrogen company was used. Subsequent data analysis is described below in *3.7.3 Cloning data analysis.*

3.7 Data analysis

3.7.1 Pyrosequencing data analysis

The obtained data from each pyrosequencing run were manually checked in PyroMark Q24 software for accuracy of the genotype calls. The β52Cys allele frequencies in the regional populations were calculated and tested for Hardy-Weinberg equilibrium in GENEPOP v 4.2 (Rousset, 2008). An exact HW test (Haldane, 1954) was used to test the hypothesis that the observed genotype frequencies correspond the expected frequencies for population in a Hardy-Weinberg equilibrium (HW). HW determines the expected genotype frequency in a population based on the observed allele frequencies. In case of two alleles the expected genotype frequencies are the expected frequencies of the homozygotes for each allele and of the heterozygotes. A deviation from HW would point to a possible erroneous genotype determination, mixing of both genes in one dataset or

that putting individuals from different sampling localities into one regional population causes significant bias in our dataset. Three variants of the exact test were used, one to test any deviation from HW in population and two others that specifically test either heterozygote excess or deficit (Rousset and Raymond, 1995).

To visualize the results, the allele frequency data was used as an input for ArcGIS v 10.2 (ESRI, Redlands, CA, USA). The allele frequency surfaces were interpolated using the Inverse distance weighted (IDW) interpolation method in 3D Analyst tools. IDW uses a method of interpolation which estimates cell values by averaging the values of sample data points in the neighbourhood of each processing cell. The IDW method was selected for two reasons. Firstly, this method does not interpolate the values outside the range of the observed values which serve as the input data. Secondly, IDW takes into account the distance of the sample data point from the center of the cell for which the value is estimated. The closer the sample data point is, the more influence or weight it has in the estimation process.

3.7.2 Sequencing data analysis

The sequences obtained by Sanger sequencing were individually aligned to a reference sequence and manually checked and corrected any errors in CodonCode Aligner (version 5.0.1.6; CodonCode Corporation, Centerville, MA, USA). As the reference sequences were used those published in GenBank by Kotlík et al. (2014). The reference sequence for HBB-T1 is accessible under the accession number KJ677170 and the one for HBB-T2 under the accession number KJ677198. All sequences for each gene separately were exported to BioEdit (version 7.0.9.0) and a complete alignment file was created. These data were used as input for DnaSP software (Librado and Rozas, 2009) in which the haplotype phase reconstruction for sequences containing two or more polymorphic sites was performed. DnaSP uses PHASE program, a coalescent-based Bayesian method (Stephens et al., 2001; Stephens and Donnelly, 2003), to infer the haplotypes.

3.7.3 Cloning data analysis

The obtained sequences were aligned to the reference sequence in CodonCode Aligner. Aligned sequences were manually checked for possible errors due to PCR recombination (Cronn et al., 2002). Such artifactual sequences were discarded and the

inferred haplotypes were exported to BioEdit. These results were compiled together with the results from the direct sequencing. These data files were used as the basic input for the subsequent phylogenetic analyses.

3.7.4 Evolutionary analysis

For the purpose of phylogenetic analyses were the BioEdit files containing the compiled data from sequencing and cloning transformed as following:

- BioEdit file for each gene separately, containing all sequences in the form of two alleles even in case of homozygous individuals
- BioEdit file containing all sequences of both genes in one alignement - for gene conversion analysis with DnaSP

3.7.4.1 Intragenic recombination

The BioEdit data files were converted to FASTA format and first analyzed by DnaSP software (Librado and Rozas, 2009). To eliminate duplicate information and lower the computational demands files with unique haplotypes were generated. Gaps and invariable sites were included.

The files with the unique haplotypes of each gene were tested for recombination which can significantly alter the phylogenetic relationships reconstruction. The HyPhy package accessible through the Datamonkey web (http://www.datamonkey.org./) (Pond and Frost, 2005; Pond and Muse, 2005) interface was used. To obtain a more comprehensive view, both Single Breakpoint Recombination (SBP) and Genetic Algorithm Recombination Detection (GARD) methods were used. SBP takes into account only one breakpoint at once, GARD on the other hand works with a hypothesis that alignment contains more breakpoints and works with these multiple breakpoints. GARD thus resolves a more complicated issue and uses a different, but potentially less accurate algorithm than SBP. Both methods in general search for segment-specific phylogenies. Sequence alignment is splitted into a left and right segment at the position of the detected breakpoint(s). Both GARD and SBP then asses whether these different segments of sequence alignment support different phylogenetic trees. Positive result points to a recombination event (Pond et al., 2006a, 2006b).

3.7.4.2 Intergenic gene conversion

The file containing the alignment of the unique haplotypes of both genes was subjected to an analysis of gene conversion. Gene conversion is important to detect for correctly inferring the phylogenetic relationships. During gene conversion, genetic material is transferred from a donor sequence, such as an allele of one gene, to the acceptor sequence, typically an allele from a paralogous gene. After the conversion event both alleles become identical in the tract that is transferred. To detect the potential gene conversion between HBB-T1 and HBB-T2 the method according to Betrán et al. (1997) which is implemented in DnaSP software was used. This test uses an algorithm to detect possible conversion tracts and then applies a statistical method to determine the number and length of the conversion tracts from the DNA sequences.

3.7.4.3 Phylogeny reconstruction

The phylogenetic trees were constructed using Mega6 software (Tamura et al., 2013) taking into the account the position of the identified break points and conversion tracts. The goal was to reconstruct the phylogenetic relationships between β52Cys and β52Ser alleles to estimate whether all β52Cys alleles are identical by descent, i.e. have the same origin. The maximum likelihood method (ML) was used for all phylogenetic tree construction. This method selects the most likely hypothesis (in our case phylogenetic tree) given the available data. The phylogenetic trees were rooted by sequences of the grey-sided vole *Clethrionomys rufocanus* that served as outgroup.

4 Results

4.1 DNA isolation

DNA was successfully extracted from 93 out of the 103 selected samples. The rest of the samples did not exhibit any DNA present on the verification electrophoresis gel or there was no PCR product amplified in the first step of pyrosequencing analysis. Interestingly, these samples have been mostly collected by external collaborators. Best performing were, as expected, spleen and liver samples, however the DNA isolation was successful even from the toe, ear and tail clips.

4.2 Pyrosequencing

Even though the recommended length of PCR amplicons to be used for pyrosequencing is only up to 500 bp, pyrosequencing assays for amplicons longer than 900 bp were successfully developed, tested and optimized.

After combining PyroMark software and OLIGO results, seven different sets of primers were prepared for testing for the major HBB-T1 gene and nine sets for the minor HBB-T2 gene. In case of HBB-T2 three assays contained the sequencing primer HBB_T1_RB_FS16 originally developed for HBB-T1 assays. This primer was the only one annealing right before the target SNP. One assay containing two amplification primers and one sequencing primer was finally selected for each gene (Tab. 1).

Tab. 1. Sequences of primers in the final selected pyrosequencing assays.

primer	sequence	type	direction
HBB_T1_RB_F19	5' GGCTGCTGGTKGTCTACCC 3'	amplification	F
HBB_T1_RB_R19	5' ACAGGCAAGTGCAGGAAAG 3'	amplification	R
HBB_T1_RB_FS16*	5' GACCTGTCCTCTGCCT 3'	sequencing	F
HBB_T2_RB_F21	5' TTGGACCCAGAGGTTCTTTGA 3'	amplification	F
HBB_T2_RB_R21	5' CTCCAGGGGACAAAAAACATT 3'	amplification	R

* The sequencing primer originally designed for HBB-T1 pyrosequencing assays was used also for HBB-T2.

A total of 518 samples out of 550 selected were successfully genotyped for both genes. Generally speaking, the samples obtained from liver or spleen gave higher peaks

than those from tail, ear or toe clips. Also samples collected by members of our laboratory were performing better that those obtained from external collaborators.

Despite the careful optimization I did not avoid occurrence of errors in some of the pyrosequencing results. In the HBB-T2 assay in 17 samples a T peak occurred at the very beginning of the sequence. Even though the peak was very low, it was a sign of an error. PyroMark dispenses one nucleotide that is not present in the sequence aways at the beginning of the sequence and before every variable position. This blank dispensation serves as a sequencing control. Using a different sequencing primer (HBB_T2_RB_FS15, for sequence see Annexe 7a) for correction runs solved the problem and the corrected sequences did not contain any unexpected peaks anymore. What was more important, genotypes on target positions obtained with the new primer differed from those obtained with the original primer HBB_T1_RB_FS16. In sequences with the T peak occurrence the target position was detected as polymorphic and the individual as heterozygous bearing both alleles. However, in the corrected sequences without the spurious T peaks all these individuals were detected as homozygous for Ser.

Seven samples amplified for HBB-T1, and four of these seven that were successfully amplified also for HBB-T2, repeatedly exhibited poor results with missing peaks within the whole sequence genotyped or with highly unreliable pattern surrounding the variable positions (for list see Annexe 7b). These samples were also rerun with a different sequencing primer and in case of HBB-T1 a completely different pyrosequencing assay with a different amlification primers was used (see Annexe 7a). However, no significant improvement was reached. Since none of the PCR exhibit any deviations that would point to non-specific amplification, these samples were not amplified nor sequenced again. They were excluded from further data analyses, but were put on the list of samples for Sanger sequencing to obtain the whole gene sequences and to reveal the possible cause of the missing peaks.

Since I was able to optimize the pyrosequencing assays for both genes to cover also the second polymorphism previously detected at the HBB-T1 gene, the Ala58Val substitution, I recorded also the results for this SNP. Val occurs at the position 58 only when there is Cys at position 52, however not always, so there is no clear association between the two amino acid substitutions. Thus no further analysis was done in this

respect. My results support the previously obtained results by Kotlík et al. (2014) with no Val being present at HBB-T2 at the position 58.

The β52Cys allele frequency was calculated for each population and tested for conformity to Hardy-Weinberg eqilibrium (HW). There were only two significant deviations from HW detected with a P value lower than 0.05, concretely in the CHEATIT population (0.04) and RO1 population (0.03). In both of these regional populations a heterozygote deficit was found. Since both of these population samples were created by merging two or more sampling sites such a result is not unexpected. This is because when samples from two or more populations with different allele frequencies are pooled, the genotype frequencies in the pool do not conform to the HW expectations even if each of the individual populations is in HW. The fact that only two out of the 70 populations showed a deviation from HW suggests that pooling sites has not seriously affected my results.

At HBB-T1, 43 out of 70 tested populations carried the β52Cys allele. Twelve of these populations were fixed for this allele and 31 polymorphic with the β52Cys allele frequency ranging from 0.06 to 0.96. At HBB-T2 the β52Cys shows rather more restricted distribution with only six populations fixed for the allele and 26 polymorphic with the β52Cys allele frequency in the range from 0.05 to 0.94 (for complete list of localities and allele frequencies see Annexe 1).

Finally, I come to the main result of the pyrosequencing analysis. Figs. 15A and B show the allele frequency surfaces interpolated in ArcGIS. They reveal a strong geographical pattern of the β52Cys allele distribution. At both genes the distribution of the β52Cys allele is largely concentrated to the west of Europe, with the majority of the fixed and high-frequency populations in France, Germany and the Netherlands. Moreover, comparison of Fig. 15A and Fig. 15C shows a clear tendency for the β52Cys allele to be associated with the Western mtDNA lineage. Surprisingly, at HBB-T1 there are two separate increases of the β52Cys allele frequency in the Balkans, concretely Bulgaria, and in southern Scandinavia, specifically Sweden (Fig. 15 A). These localities belong to the Balkan and Carpathian mtDNA lineage, respectively (Fig. 15 C).

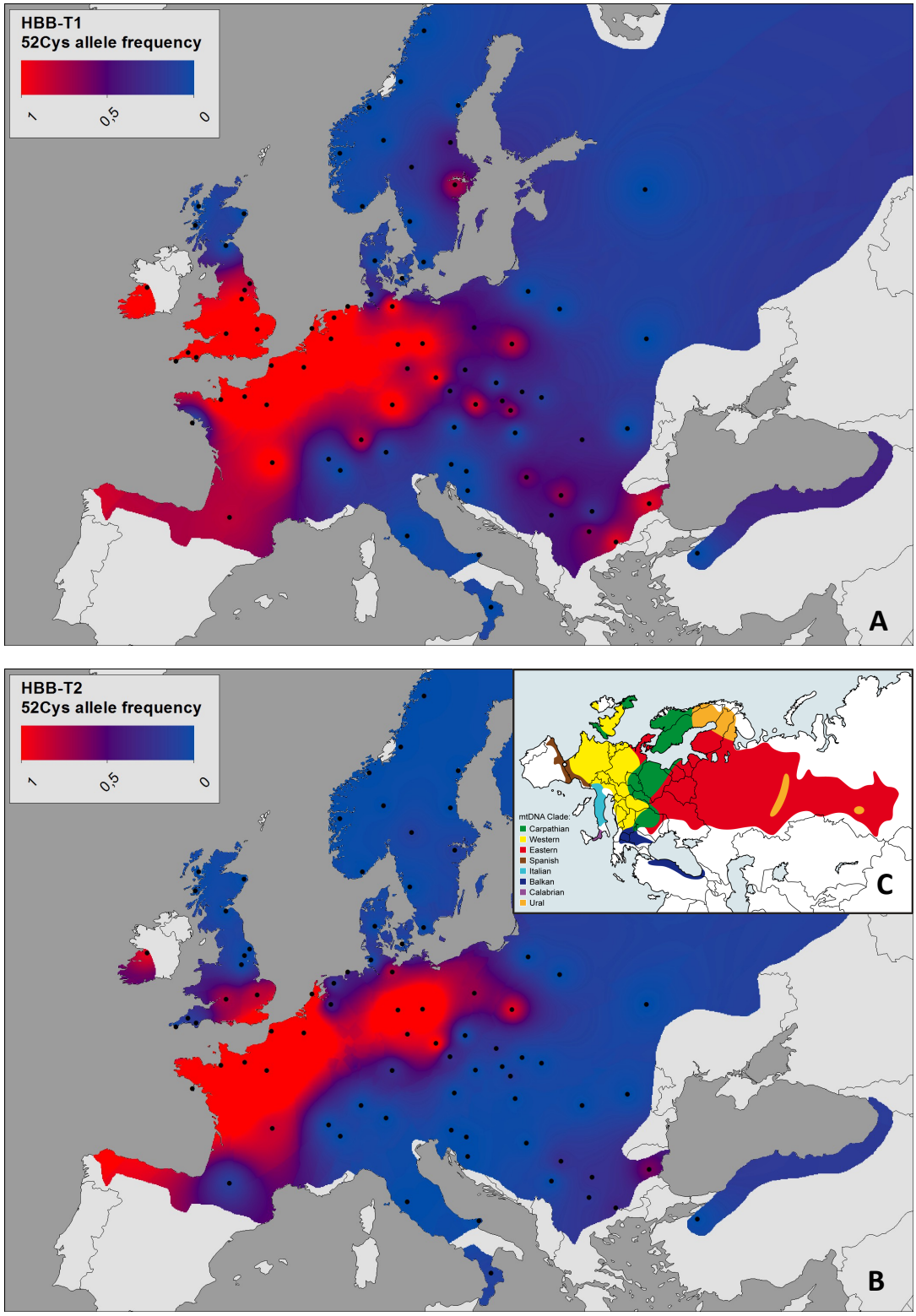


Fig. 15 Geographic distribution of β52Cys allele at HBB-T1 (**A**) and HBB-T2 (**B**) shown as interpolated allele frequency surfaces. Dots represent centres of population samples. Data for Britain were taken from Kotlík et al. (2014). **C** represents mtDNA lineages distribution as revised by Filipi et al. (2015).

4.3 Sanger sequencing

At HBB-T1, for 21 out of 30 analyzed samples sequence data were obtained successfully by direct sequencing. The remaining 9 were discarded due to repeatedly poor performance in PCR amplification step. Sequence data were obtained also from the seven samples added to the list due to the missing peaks they exhibited in the pyrosequencing analysis. A total amount of eight additional bank vole sequences available from previous study was added to the dataset. However, for 9 sequenced heterozygous samples I was unable to resolve the phase of the haplotypes by PHASE software and had to use cloning. Five out of these nine samples contained indels in their sequences.

At HBB-T2, 21 out of the 28 analyzed samples were successfully sequenced as well as the four out of the seven samples showing poor pyrosequencing results. Again, the remaining 10 samples were discarded due to poor PCR amplification performance. A total amount of 10 additional bank vole sequences available from previous study was added to the dataset. A total of 10 heterozygous samples were cloned to obtain phased haplotypes. For complete list of samples see Annexe 5.

4.4 Cloning

Cloning was successful for all samples subjected to it. Nevertheless, it had to be repeated for three samples. These either produced only low number of recombinant colonies that could be used for sequencing and/or the sequences obtained from these samples were of very poor quality, often containing high number of PCR artifacts. These samples were cloned again. To minimize the probability of recombinant sequences formation during PCR, the PCR programme used for the amplification was modified. The amplification step was prolonged to avoid not fully amplified sequences to be produced and the number of cycles was lowered (for the exact PCR programme see Annexe XY). To assure the best possible performance, I also used a fresh stock of the pDrive vector from a newer cloning kit.

The sequences obtained by cloning were aligned with the sequences obtained by direct Sanger sequencing of the PCR products and included in the final dataset. An outgroup sequence of *C. rufocanus* already obtained by Sanger sequencing for the purpose of a previous study was added to the dataset. After editing the files so that each

sample was represented by two sequences, either identical or different depending on whether it was a homozygote or heterozygote, the resulting files contained 74 sequences for HBB-T1 and 72 sequences for HBB-T2, with the difference being due to differential amplification success between the genes in some individuals.

4.5 Evolutionary analysis

4.5.1 Recombination

DnaSP produced 38 unique haplotypes from the 74 HBB-T1 sequences and 43 haplotypes from the 72 HBB-T2 sequences. The file with all the haplotypes from both genes aligned together thus contains 81 unique haplotypes. (For the complete list of haplotypes, individuals carrying them and the sampling localities, see Annexe 8) This number is including the outgroup *C. rufocanus* sequences that account for 3 haplotypes in total. The outgroup haplotypes were omitted from the recombination and gene conversion analysis.

The SBP analysis identified significant breakpoints in both the HBB-T1 and HBB-T2 datasets (Tab. 2). For each gene, the most probable breakpoint was detected on same position with both the AIC and cAIC criterion (Tab. 2). The GARD analysis did not identify any significant breakpoint according to AIC or cAIC.

Tab. 2. Breakpoints with strong support identified by SBP analysis in DataMonkey. The most probable breakpoint at each gene is highlighted.

	breakpoint	AIC	AIC support	cAIC	cAIC support
HBB-T1 haplotypes	514	6047.01	0.22843	6083.56	0.225731
	521	6046.09	0.362554	6082.63	0.35827
HBB-T2 haplotypes	476	4753.39	0.687908	4809.78	0.518795

4.5.2 Gene conversion

A significant presence of conversion tracts was detected in haplotypes from both HBB-T1 and HBB-T2. In total, 12 different conversion tracts were identified affecting 23 haplotypes (some tracts were present in multiple haplotypes, for figure see Annexe 9). Three of these conversion tracts, affecting six haplotypes, included the target SNP position Ser52Cys. Only one of these tracts affects HBB-T1 haplotypes, but three of them. These are one haplotype from France (Hap_18) and two from Sweden (Hap_21 and Hap_22),

from the locality where the separate increase of the β52Cys allele frequency was observed (Fig. 15A). In two of these three HBB-T1 haplotypes (Hap_18 and Hap22), the conversion tract contained a codon for Cys and in the third haplotype (Hap_21) a codon for Ser. The SBP analysis identified possible breakpoints in both the HBB-T1 and HBB-T2 datasets.

4.5.3 Phylogenetic relationships

In Figs. 16 through 19 there are phylogenetic trees reconstructed by ML and showing the relationships among the HBB-T1 and HBB-T2 haplotypes, separately for each gene, and taking into account the results of the recombination analysis. Thus, two haplotype trees are presented for each gene, one representing the gene segment to the left of the breakpoint and including it (Figs. 16 and 18) and the other the segment to the right of the breakpoint (Figs. 17 and 19). For any reference of the haplotypes and origin of samples included in them see Annexe XY.

Fig. 16 represents the haplotype tree for the HBB-T1 gene, specifically for the gene segment to the left side of the breakpoint located at the site 521. To estimate the tree the Kimura 2-parameter substitution model (K2) with an allowance for the presence of invariant sites (+I) was used, which was found to be the best-fit model for this part of the alignment. The tree shows that the haplotypes containing the 52Cys allele cluster to three distinct groups hereafter referred to as ‘Eastern’, ‘Western’ and ‘SWE+FR’. The ‘Eastern’ haplogroup is comprised of 5 haplotypes from eastern Europe, concretely Hap 2 (BGR), Hap 6 (SRB), Hap 11 (POL + SVK), Hap 12 (POL + SVK) and Hap 14 (POL). The ‘Western’ group consists of 4 haplotypes from the Czech Republic and western Europe, specifically Hap 7 (CZE), Hap 8 (CZE), Hap 10 (DEU) and Hap 20 (FRA). The ‘SWE+FR’ haplogroup is, as the name implies, formed by haplotypes from France (Hap 18) and Sweden (Hap 22).

The tree in Fig. 17, representing the HBB-T1 gene segment to the right side of the breakpoint located at the site 521, was constructed using the K2 model with the evolutionary rate variation among sites modeled by a Gamma distribution (+G). This tree is less well resolved, but the previously defined ‘SWE+FR’ and ‘Western’ haplogroups are retained. The ‘Eastern’ group is retained only partly, since the Hap 2 (BGR) now clusters with haplotypes of Calabrian (Hap 26 and Hap 28) and Italian mtDNA lineage (Hap 34).

Fig. 18 represents the tree for the HBB-T2 gene, again the left side of the gene segment including the breakpoint located at the site 476. It was constructed using the Hasegawa, Kishino and Yano model (HKY) with an allowance for invariant sites (+I). In this tree the 52Cys haplotypes form only two distinct groups. Five haplotypes from both eastern and western Europe, specifically Hap 46 (CZE), Hap 50 (POL), Hap 52 (FRA), Hap 54 (FRA) and Hap 60 (DEU), form one of the groups, referred to as 'East+West' in Fig. 18. The second group is consisting of two haplotypes Hap 53 (FRA) and Hap 56 (SWE + FRA + SVK).

The tree in Fig. 19 representing the right side HBB-T2 gene segment was constructed using the K2+G model. In contrary to the HBB-T1 gene where the groups were more or less present in both gene segments, this tree demonstrates that in HBB-T2 are the groups defined in the left segment disintegrated completely in the tree based of the right segment (Fig. 19).

The last two figures (Figs. 20 and 21) show phylogenetic trees for the haplotypes of both genes together, but excluding all haplotypes containing the conversion tracts except for the three HBB-T1 haplotypes Hap_18, Hap_21 and Hap_22. The conversion tract present in these three haplotypes contained the polymorphic site Ser52Cys and the trees show the phylogenetic relationships for the gene segment corresponding to this conversion tract (sites 208-664; Fig. 20) and for the remaining two segments of the gene (concatenated sites 1-207 and 665-1128; Fig. 21). Both trees were constructed by the ML method using the K2+G substitution model.

The tree in Fig. 20 representing the gene segment corresponding to the conversion tract, shows that the HBB-T1 52Cys haplotypes from the 'SWE+FR' group cluster with the HBB-T2 52Cys haplotypes with high statistical bootstrap support. The third HBB-T1 haplotype clustering with the HBB-T2 haplotypes is the 52Ser haplotype from Sweden (Hap 21). This haplotype was also detected as containing the same converted tract, but with the 52Ser allele (Annexe 9).

In contrary, the tree in Fig. 21 representing the phylogeny of the two remaining segments of the gene shows that all the haplotypes are clustered according the gene when the converted tract is excluded.

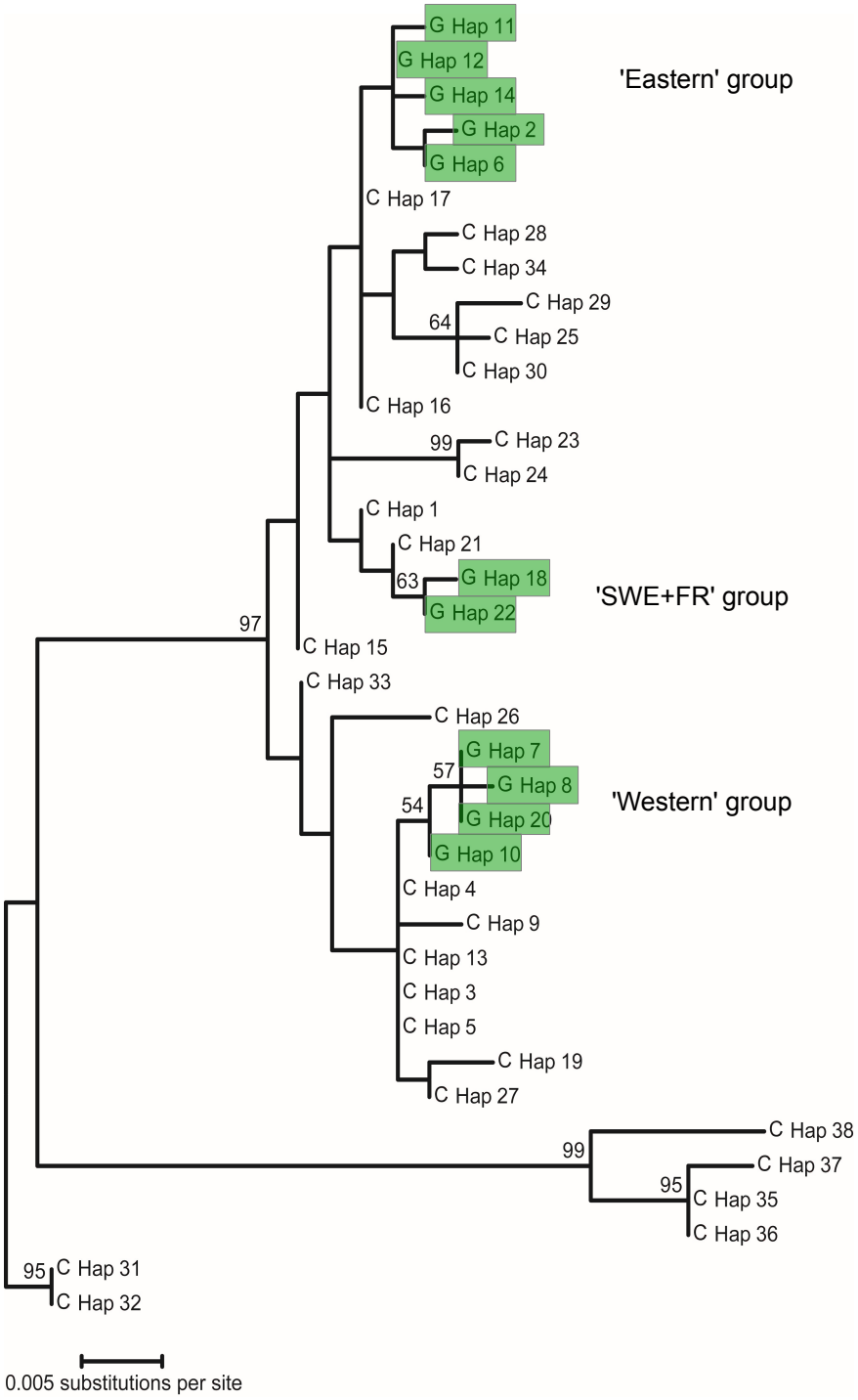


Fig. 16 Maximum likelihood phylogeny of HBB-T1 haplotypes based on the alignment segment left of the breakpoint at site 521. C and G represent Ser and Cys at site 52, respectively. Haplotypes containing Cys at site 52 are highlighted. Percentage bootstrap support for branches based on 1000 replicates is shown when higher than 50%. Major groups of 52Cys containing haplotypes are labelled at right. Hap 31 and 32 are haplotypes of *C.rufocanus* serving as outgroup.

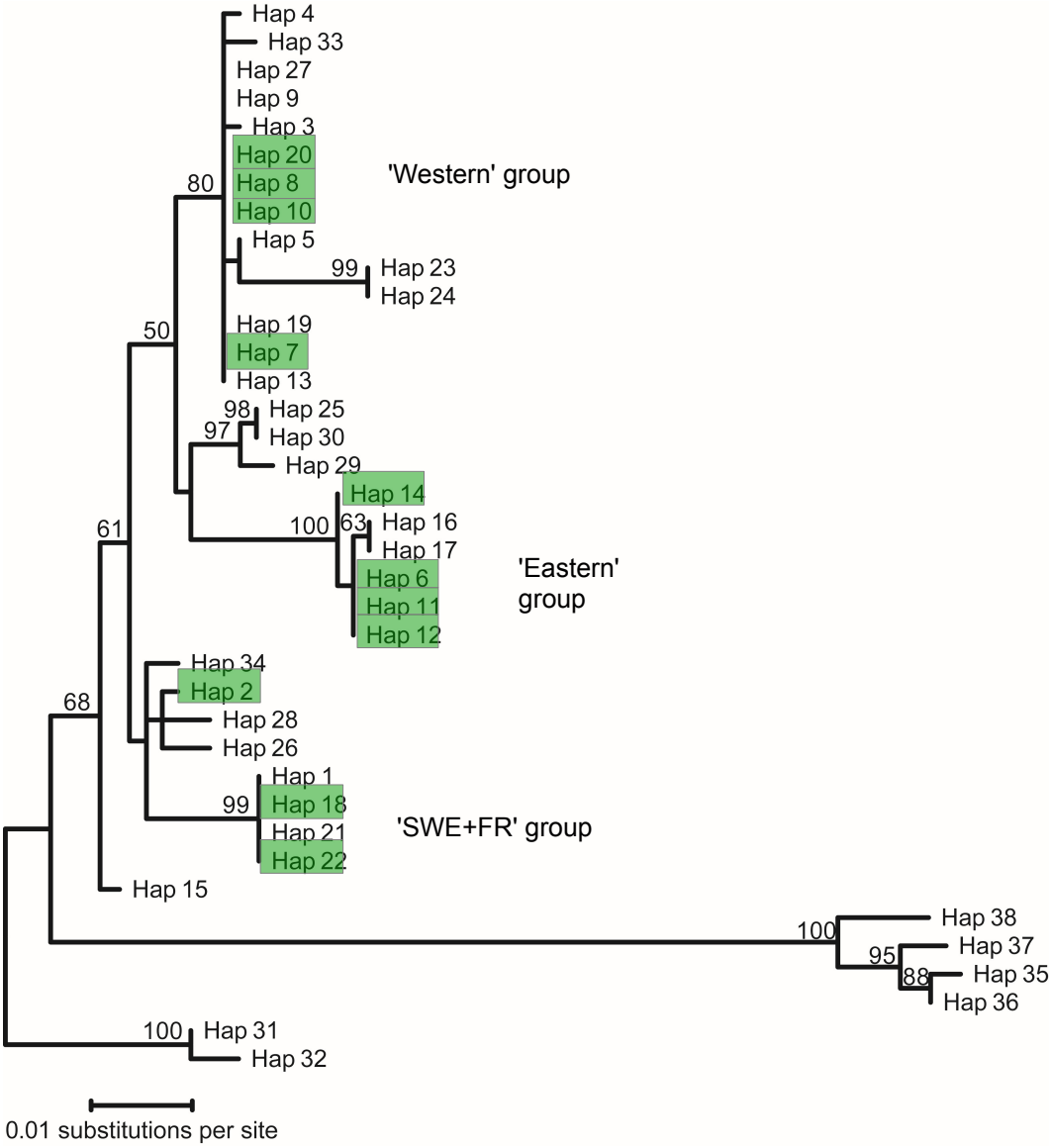


Fig. 17 Maximum likelihood phylogeny of HBB-T1 haplotypes based on the alignment segment right of the breakpoint at site 521. Haplotypes containing Cys at site 52 are highlighted. Percentage bootstrap support for branches based on 1000 replicates is shown when higher than 50%. Major groups of 52Cys containing haplotypes are labelled at right. Hap 31 and 32 are haplotypes of *C.rufocanus* serving as outgroup.

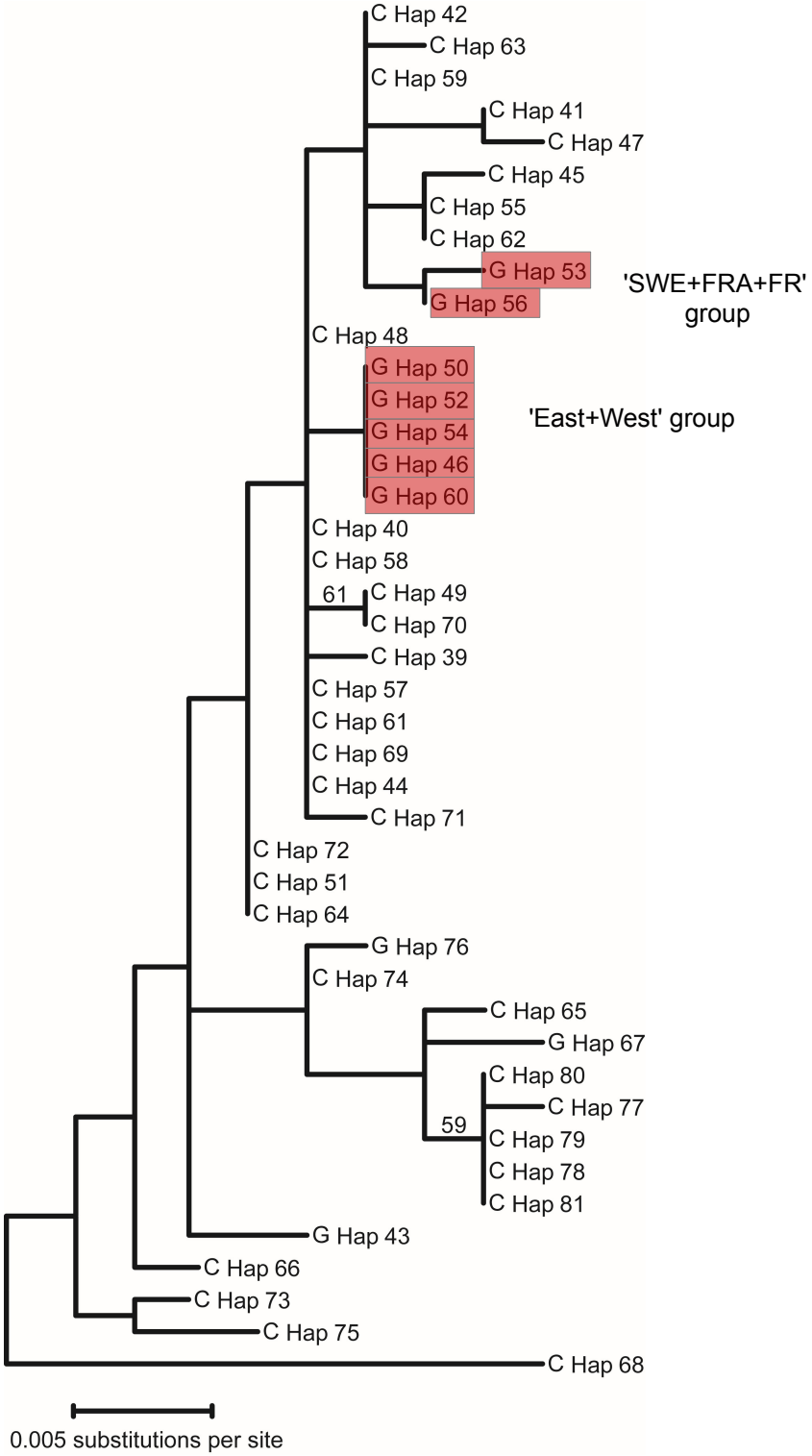


Fig. 18 Maximum likelihood phylogeny of HBB-T2 haplotypes based on the alignment segment left of the breakpoint at site 476. C and G represent Ser and Cys at site 52, respectively. Haplotypes containing Cys at site 52 are highlighted. Percentage bootstrap support for branches based on 1000 replicates is shown when higher than 50%. Major groups of 52Cys containing haplotypes are labelled at right. Hap 68 is haplotype of *C.rufocanus* serving as outgroup.

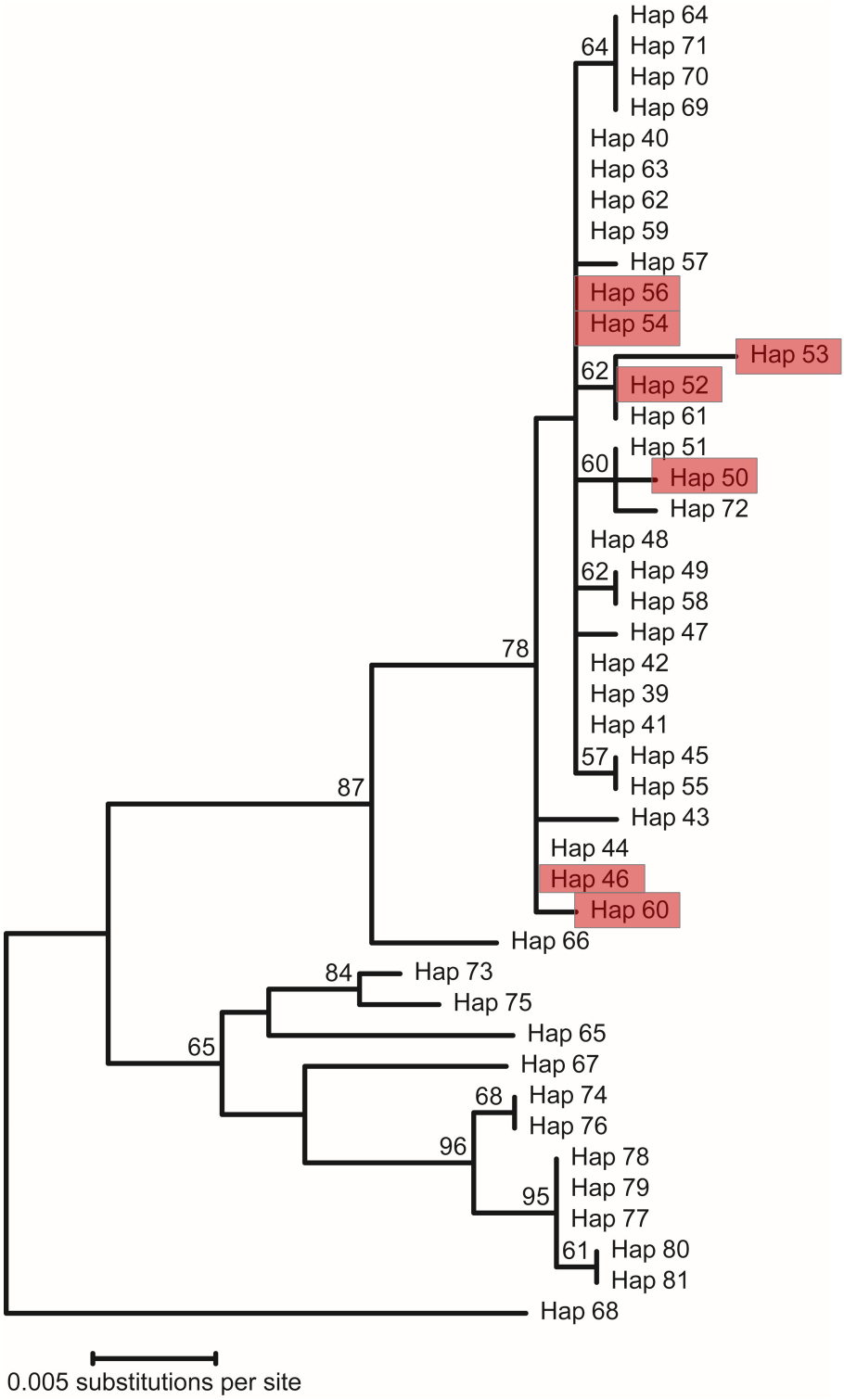


Fig. 19 Maximum likelihood phylogeny of HBB-T2 haplotypes based on the alignment segment right of the breakpoint at site 476. Haplotypes containing Cys at site 52 are highlighted. Percentage bootstrap support for branches based on 1000 replicates is shown when higher than 50%. Hap 68 is haplotype of *C.rufocanus* serving as outgroup.

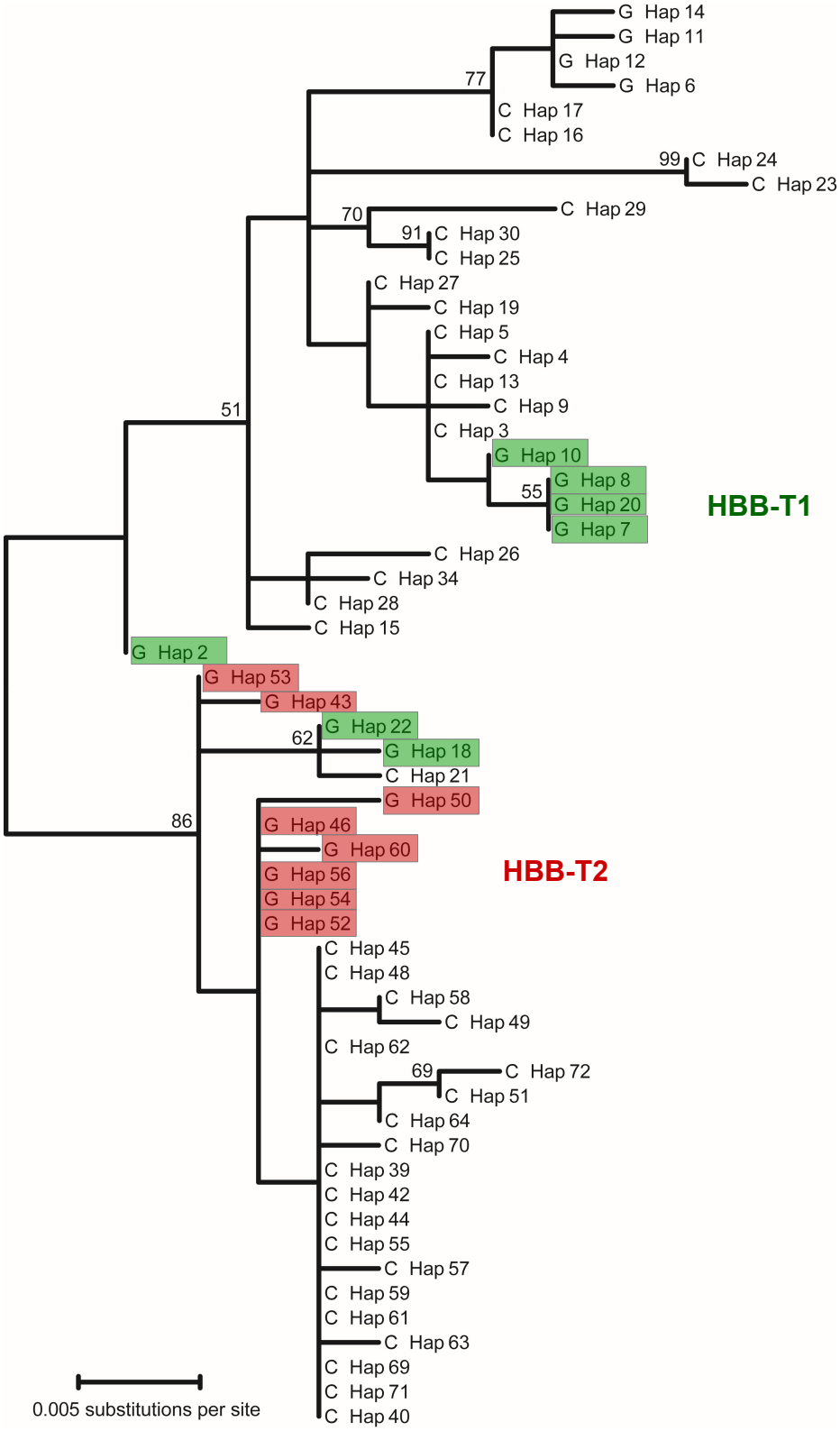


Fig. 20 Maximum likelihood phylogeny for both genes, representing the gene segment corresponding to the conversion tract identified in Hap 18, 21 and 22 between the sites 208-664. Other haplotypes containing conversion tracts are excluded. C and G letters represent 52Ser and 52Cys allele, respectively. Haplotypes containing Cys at site 52 are highlighted according to the genes. Percentage bootstrap support for branches based on 1000 replicates is shown when higher than 50%.

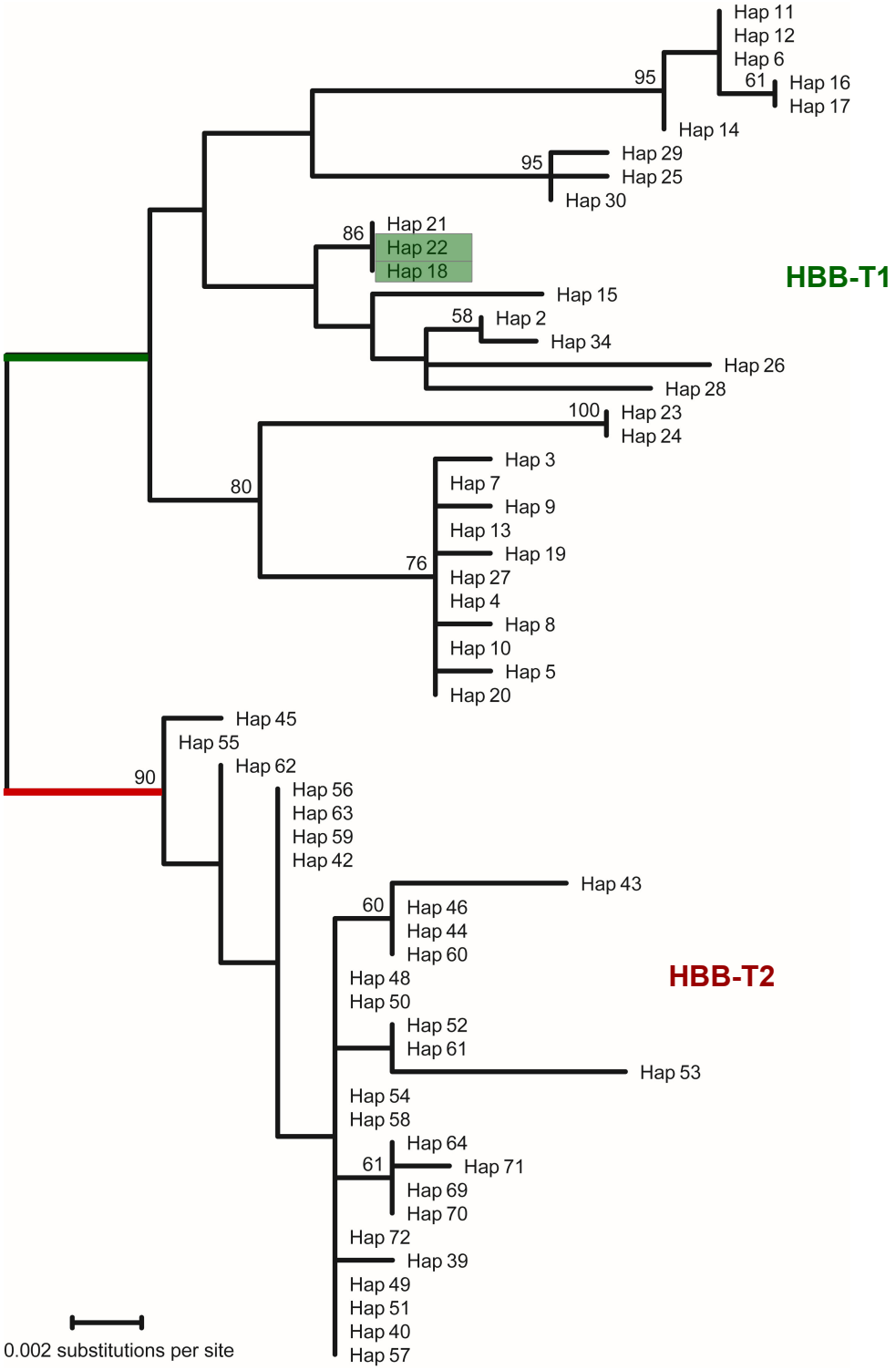


Fig. 21 Maximum likelihood phylogeny for both genes, representing the remaining two segments of the genes (concatenated sites 1-207 and 665-1128) after exclusion of the segment corresponding to the converted tract. Other haplotypes containing conversion tracts were excluded. Haplotypes 18 and 21 containing Cys in the conversion tract are highlighted. Percentage bootstrap support for branches based on 1000 replicates is shown when higher than 50%.

5 Discussion

5.1 Technical aspects of work

Regarding the technical aspects of my work, the difficulties during the DNA isolation and pyrosequencing that I faced with the samples collected by external collaborators confirm that the method of collection and preservation of the samples plays a very important role for the result quality of subsequent analyses. The samples collected by the members of our laboratory were taken mostly from live trapped animals and the tissues collected were immediately fixed in 96% ethanol. Even though I don't have the information about how the samples donated by external collaborators were collected, preserved and overall handled, I argue that these procedures were not carried out with sufficient caution and precision and the DNA in these samples was degraded.

The pyrosequencing performed on the PyroMark Q24 platform from Qiagen proved to be an effective and reliable method for genotyping of SNP within a short segment of DNA sequence. Regarding my research, as the most significant advantages I see for example the easy sample preparation with available commercial kits, speed of the pyrosequencing analysis when results for 24 samples are available in approximately 20 minutes and also the possible control of genotype calls accuracy. However, to make the pyrosequencing method fast and cost-effective, it is necessary to spend some time with optimization prior to the onset of a massive sample analysis. This is especially true when analyzing a PCR product of nonstandart length as in my case.

5.2 Pyrosequencing assay optimization and challenges encountered

Despite the initial worries that the PCR product with the length over 900 bp would be too long for a pyrosequencing analysis and would for example break down during the purification, I was able to successfully develop and optimize assays for this longer PCR product. Thanks to the protocols provided by the Qiagen company the optimization process was relatively fast and easy. The clearly defined conditions and procedure of the sample preparation simplify identification of the factors potentially responsible for any suboptimal results. Based on my own work I can draw the general conclusion that the quality of the outcome from the pyrosequencing analysis of long PCR products is strongly influenced by the amount of the streptavidin-coated sepharose beads used in the

immobilization step. The optimization process also confirmed that the best approach is to design the sequencing primer as close to the target position as possible since especially in case of DNA of suboptimal quality the accuracy of the read is lowering very quickly with increasing distance from the sequencing primer.

As it was already described in the results, even the most careful optimization did not lead to complete suppression of sequencing artifacts. In some samples genotyped for the HBB-T2 gene an unexpected T peak occurred in the very beginning of the read. The sequencing primer was designed to terminate exactly before the target position and thus the first incorporated base should have been C or G. The T base was dispensed by PyroMark only as a blank sequencing control and as such, it should not have been incorporated in the sequence and no peak should have been recorded. However, a T base is in fact present in the sequence at the position prior the target SNP, i.e. it is the last base of the sequencing primer. The occurrence of the extra T peak can be thus most easily explained by a possible degradation of the sequencing primer from its 3' end. This explanation seems to be highly probable since the prepared primers were repeatedly thawed and frozen.

The missing peaks in the samples from Norway, Sweden, France and Switzerland listed in Annexe XY seemed at first to be more complicated to explain. Using a different set of amplification and sequencing primers did not solve the problem. The fact that two different sets of primers resulted in exactly the same error makes potential explanations such as mispriming, unspecific annealing of the primers, and a hairpin loop or dimer formation not very likely. Comparing the primer sequences with the results from the Sanger sequencing revealed that all amplification and sequencing primers matched perfectly with all the error exhibiting sequences. Nevertheless, after closer examination of the results I can provide a rather simple explanation for the missing peaks. At HBB-T1 the samples 1670, 1859, 1871 and 1879 exhibited the missing peaks specifically after the second variable position responsible for the Ala58Val aminoacid substitution. From the results of the Sanger sequencing, I found out that these sequences have the Ala coded by a GCT triplet instead of the prevailing GCC. Similarily, the 1048, 1395 and 1911 samples that exhibited the missing peaks along the whole sequence determined by pyrosequencing have the 52Ser aminoacid coded by TCT instead of TCC. At HBB-T2 I have

no data for the samples 1048, 1395 and 1911 due to the poor amplification performance, nevertheless the rest of the samples, i.e. 1670, 1859, 1871 and 1879, exhibited missing peaks along the whole sequence and have the 52Ser coded again by a TCT triplet instead of TCC. These single nucleotide substitutions do not change the amino acid sequence, however, they are unexpected and were not taken into the account in the design of the pyrosequencing assays. Thus, the dispensation order did not match the template DNA sequence and the erroneous results were produced. The pyrosequencing assays will be probably modified according to these findings before the onset of further analysis on a large scale, even though such substitutions are a minority.

5.3 β52Cys allele distribution and origin

One of my main goals was to map the β52Cys allele distribution in the continental Europe and to determine its most likely geographic origin. By means of the allele frequency surface interpolation I revealed a strong geographical pattern of the β52Cys allele distribution. Populations fixed for the β52Cys allele or with a high β52Cys allele frequency are concentrated to western Europe, specifically to France, Germany and the Netherlands. Moreover, when comparing the β52Cys allele distribution at the HBB-T1 gene with the distribution of the bank vole mtDNA lineages as updated by Filipi et al. (2015), there is an evident tendency of the β52Cys allele to be associated with the Western mtDNA lineage (see Fig. 15 in *4. Results*).

This finding importantly extends the current knowledge of the β52Cys allele distribution until now only mapped in Britain by Kotlík et al. (2014). Based on preliminary mtDNA results, the individuals in Britain bearing the 52Cys haemoglobin variant were first tentatively linked to the continental populations in the Netherlands also carrying 52Cys (Kotlík et al., 2014), which was later confirmed by showing that the first British colonists belonged the Western mtDNA lineage (Filipi et al., 2015). My results show that this link between the 52Cys haemoglobin and the western mtDNA lineage exists also in other populations in the continental Europe. Since the Westen lineage most likely originated from woodland refugia in the foothillls of the Alps (Kotlík et al., 2006; Magri et al., 2006), I hypothesize that these refugia were also the most likely geographical origin of the β52Cys allele and that it therefore dispersed post-glacially with the bank voles of the Western

mtDNA lineage. Nevertheless, there are two interesting features arising from my datasets that fall outside the pattern of the correlation with the Western mtDNA lineage, and thus deserve a closer inspection.

Firstly, the β52Cys allele is absent from the populations of the Western mtDNA lineage within the mountain range of the Alps in central Europe and from the Dinaric Mountains in the Balkans. This tendency of the β52Cys allele to be concentrated to the areas of lower altitude in the continental Europe correlates with the β52Cys allele distribution centered mostly in southern lowland areas of Britain as described by Kotlík et al. (2014). My findings from the continental Europe only emphasize that such a clearly defined distribution is not likely to be maintained without an involvement of selection. The delimited distribution of the β52Cys allele means that Cys might be advantageous, but not universally, only under certain conditions. Bearing in mind the higher antioxidant capacity of red blood cells of bank voles possessing the 52Cys haemoglobin variant, my results support the hypothesis that the 52Cys haemoglobin variant could have been adaptive advantage under certain climatic conditions for its bearers during the postglacial colonization. While in Britain this adaptive advantage probably promoted the population replacement of the after LGM (Kotlík et al., 2014), my findings show that it could have been similarly important in shaping the current geographical distribution of the bank vole in the continental Europe. The hypothesis that the first colonists of Britain bearing 52Ser were better suited for colder climatic conditions and then during warming they were outcompeted by the second colonists bearing 52Cys is applicable in certain extent also for the continental populations. The discontinuous distributions of some mtDNA lineages (e.g. Carpathian and Eastern) suggest a rather complex history of colonization and replacement (Fig. 2) (Filipi et al.; 2015). It is possible that the individuals bearing the 52Cys haemoglobin could afford producing a higher number of offspring than the individuals with 52Ser, but without the burden of higher levels of oxidative damage to their tissues. This quality could have manifested as advantageous in the lowland areas with milder climate where individuals are most probably capable of higher energy intake and where they are allowed to put more energy into reproduction thanks to lower demands on thermoregulation. Thus the β52Cys allele could have been an adaptive advantage that lead to a more rapid population expansion and establishment in such areas with favorable

climatic and environmental conditions. Such a genetically based predisposition to a higher reproductive rate conferred by the β52Cys haemoglobin may have been difficult to outcompete for the members of the populations with the 52Ser variant.

Second feature that deviates from the pattern of association with the Western mtDNA lineage are the two unexpected separate increases of the β52Cys allele frequency in Sweden and in the Balkan region. These populations do not belong to the Western mtDNA lineage and this fact significantly challenges the idea of the single origin of the 52Cys allele. The Ser52Cys amino acid substitution can be found, albeit to a lesser extent, also at the minor HBB-T2 gene, which shows much lower expression level in the bank vole than HBB-T1, and is thus also less functionally significant. This leads to a tempting idea of a role of gene conversion between HBB-T1 and HBB-T2 as a means of an intergenic transfer of the 52Cys allele. The single origin of the 52Cys allele together with this hypothesis about gene conversion was tested by subsequent phylogenetic analysis.

5.4 Phylogenetic relationships and the role of gene conversion

Prior to any phylogenetic analysis it was necessary to analyze my data to detect signs of possible recombination and gene conversion. Recombination can significantly bias the interpretation of the phylogenetic relationships if not detected and taken into account. As an excellent example can serve the Hap 2 of HBB-T1 gene which clusters with the eastern group of haplotypes in one tree (Fig. 16) and with the basal Calabrian haplotypes in the second tree (Fig. 17), depending which part of the alignment (left or right) divided by recombination into two parts the phylogenetic tree was constructed for.

The phylogeny of both HBB-T1 and HBB-T2 genes exhibits significant signs of recombination. However, in both genes, the phylogenetic trees representing the part of the alignment to the left from the detected breakpoint, which includes our target position, the 52Cys alleles form three and two distinct groups, respectively (Fig. 16 and 18, respectively).

Since HBB-T1 clearly is functionally more significant than HBB-T2 gene due to the much higher expression level, I will focus the on the results regarding the major gene HBB-T1. The existence of three distinct clusters of HBB-T1 haplotypes carrying the 52Cys allele (Fig. 16) points to a fact that the 52Cys alleles from the different parts of the bank

vole distribution are not identical by descent, i.e. they don't have a single mutational origin. This is rather unequivocally supported by the cluster containing Hap 18 and Hap 22, which are the result of a gene conversion as will be discussed later. The other two clusters of the “Eastern” and “Western” haplotypes (comprised of Hap 2, 6, 11, 12, 14 and Hap 7, 8, 10, 20, respectively) could mean there have been parallel mutations at the same site that were picked by positive selection as independent random increases in frequency by genetic drift would appear unlikely. Alternatively, the two clusters can be due to an undetected recombination or conversion event. There are no significant signatures of recombination or gene conversion that would explain it. However, such events are detectable only when the alignments contain enough phylogenetic signal supporting them, such as unbroken tracts of multiple nucleotides in some haplotypes that match the sequences of another haplotypes of the same (in the case of intragenic recombination) or different gene (in the case of intergenic gene conversion). Unfortunately, as time passes the transferred sequence blocks get broken down by subsequent recombination events, eventually leading to a loss of the signal. Thus I can not exclude an undetected recombination/conversion as the agent responsible for the clustering of the 52Cys alleles into the two groups.

Despite this uncertainity about the mechanism underlying the existence of the Eastern and Western haplotypes clusters, the existence of Hap 18 and Hap 22 in a separate cluster can only be explained by gene conversion, which supports the hypothesis about the multiple origins for the presence of 52Cys at the functionally important HBB-T1 gene. Gene conversion tracts have been identified in both HBB-T1 and HBB-T2. However, only a single tract contained the Ser52Cys substitution and it affected three HBB-T1 haplotypes, specifically Hap 18 and Hap 22 with the 52Cys allele and Hap 21 with the 52Ser allele. In fact, the presence of both alleles suggests the tracts in Hap 18 and Hap 22 and in Hap 21 may stem from different conversion event, even though they span the same nucleotides in all three haplotypes. This conversion tract was handled separately in the phylogenetic analysis where the nonconverted parts of the three haplotypes were clustered with the remaining HBB-T1 haplotypes while their sequences in the converted tract were significantly clustered with the HBB-T2 haplotypes. This confirms the conversion event from HBB-T2 to HBB-T1 and the fact that this conversion tract contained

the target SNP Ser52Cys supports the hypothesis, that the gene conversion can be a means by which the 52Cys allele is transferred between the genes.

The fact that the only conversion event detected that included 52Cys was the conversion in the direction from HBB-T2 to HBB-T1 does not necessarily mean that the mutation Ser52Cys originated in HBB-T2 and was then converted into HBB-T1. However, regardless the gene of origin of the 52Cys allele this finding does have important implications about the functional evolution of the bank vole haemoglobin. This is because it suggests that the low expressed HBB-T2 gene can serve as a reservoir and source of the functional variation at the site β52 for HBB-T1. It is likely that due to the high reactivity of its thiol the 52Cys allele is not advantageous universally, but only when there is an increased need for antioxidants. The presence of the 52Cys allele in the minor gene would not significantly affect the functionality of haemoglobin due to the low expression of HBB-T2 and thus it would not be under strong selection pressure there, positive or negative. Nevertheless, this way maintained in the population the 52Cys allele could once in a while be transferred by gene conversion from HBB-T2 into the highly expressed HBB-T1 gene and there tested by natural selection. In the absence of favourable conditions (i.e. of strong selection pressure favouring red blood cells with increased antioxidant capacity), the converted HBB-T1 haplotypes containing 52Cys would be eliminated from the population by random drift or negative selection, depending on whether 52Cys would give some disadvantage to its bearers. However, in a situation of increased need for antioxidants, let say under unfavourable climatic, population density or epidemiological conditions, a positive selection could pick the 52Cys allele and rapidly increase its frequency possibly up to fixation, as it would likely confer an advantage over the individuals bearing the 52Ser allele. This could potentially have been the case of the Swedish populations where Hap 18 and Hap 22 were exclusively found.

An interesting feature also emerged regarding the samples 1048, 1395 an 1911 which have the 52Ser at HBB-T1 coded by the TCT triplet instead of TCC. Interestingly, in the phylogenetic trees these haplotypes cluster into a separate clade that appears more divergent from the rest of the bank voles HBB-T1 haplotypes than are the haplotypes of the outgroup *C. rufocanus*. I examined these haplotypes more closely and I found that while they are all homozygous for 52Ser they posses four other non-synonymous amino

acid substitutions that distinguish them from all the other HBB-T1 haplotypes. These are Val18Ala, Gly22Ala, Thr69His and Gly72Ala substitutions. Two of these, Gly22Ala and Thr69His are also present in the outgroup haplotypes of *C. rufocanus*. The other two are not found in any other sequences within my dataset. It is possible that these particular haplotypes are a result of an undetected gene conversion event. It is also possible, that the haplotypes are relicts of some ancient lineage surviving in the areas of origin of the samples and we sampled these by chance.

5.5 Future prospects

My results map the current distribution of the functionally important β52Cys allele of the bank vole haemoglobin in the continental Europe and determine its possible geographical origin. The subsequent phylogenetic analysis revealed that the 52Cys alleles from the different parts of the distribution area necessarily do not have only one single evolutionary origin. Furthermore, it proved the importance of gene conversion as a means of transfer of the 52Cys allele between the two genes, which would change its exposure to selection pressure. All these findings help to understand the postglacial colonization of Europe by temperate species and emphasize the importance of intraspecific variability and selection for surviving a climatic change. Nevertheless, even though the current distribution of adaptive variation can give as an important clue about how the postglacial colonization took place, it still reflects mostly the adaptations of the populations to the current climatic conditions in the areas they occupy at present. Thus to reveal more precisely how the processes in the past impacted the distribution during the different stages of recolonization it would be necessary to involve also a combination of paleoclimatic data, ecological niche modelling into the past and fossil and ancient DNA data where available. These information when carefully interpreted could give us more comprehensive view on how exactly the distribution of the 52Cys allele, and of the populations carrying it, changed in time under the selection pressure imposed by climatic change. Together with that, it is important to bear in mind, that even though the haemoglobin genes exhibit the variability that seems to be an adaptive advantage under certain climatic conditions, so far it could not be directly linked with climate adaptation. Haemoglobin is only a part of a very complex physiological network and it is very likely

that there are other important genes involved in the coping of bank voles with a climatic change. Thus, it would be interesting to search for other candidate SNPs within the bank vole genome and analyze their possible functional significance as it was done with the SNP responsible for two haemoglobin variants by Kotlík et al. (2014).

6 Conclusion

Previous studies of the bank vole adaptive phylogeography in Britain documented that the intraspecific functional haemoglobin variability and selection acting upon it may have promoted the population replacement in this species during the postglacial recolonization of Britain. My study builds on these findings and suggests that the genetic variability at both the major and minor β-globin genes among the different continental bank vole populations was of similar importance for shaping their current distributions. Primarily, the results of my study revealed the importance of the gene conversion as a means of transfer of the responsible functional mutation between the major and the minor gene. Since the minor gene is approximately 24 fold less expressed, it is also under much less selection pressure than the major gene. Since the mutation likely is not universally advantageous, the presence of the mutation in the minor gene can help to maintain it in the population during conditions that do not favour it. Transfer of the mutation into the major gene by gene conversion would expose it to strong selection pressure and either would lead to elimination of the mutation from the major gene or to its fixation in the population, if it manifests as advantageous. The importance of this finding lays in the fact, that it shows that even though different populations of the same species may seem to be uniform at first sight, their genetic constitution may vary significantly. Moreover, since this intraspecific variability may be of functional significance also the importance of the populations for species survival under specific climatic conditions may vary.

7 References

Ahmadian, A., Ehn, M., Hober, S., 2006. Pyrosequencing: history, biochemistry and future. Clinica chimica acta 363, 83–94.

Ahmadian, A., Gharizadeh, B., Gustafsson, A.C., Sterky, F., Nyrén, P., Uhlén, M., Lundeberg, J., 2000. Single-Nucleotide Polymorphism Analysis by Pyrosequencing. Analytical Biochemistry 280, 103–110. doi:10.1006/abio.2000.4493

Avise, J.C., 2009. Phylogeography: retrospect and prospect. Journal of biogeography 36, 3–15.

Avise, J.C., 1998. The history and purview of phylogeography: a personal reflection. Molecular Ecology 7, 371–379.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., Saunders, N.C., 1987. Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. Annual Review of Ecology and Systematics 18, 489–522.

Barnes, I., Matheus, P., Shapiro, B., Jensen, D., Cooper, A., 2002. Dynamics of Pleistocene population extinctions in Beringian brown bears. Science 295, 2267–2270.

Barton, R.N.E., Jacobi, R.M., Stapert, D., Street, M.J., 2003. The Late-glacial reoccupation of the British Isles and the Creswellian. J. Quaternary Sci. 18, 631–643. doi:10.1002/jqs.772

Bazin, E., Glémin, S., Galtier, N., 2006. Population Size Does Not Influence Mitochondrial Genetic Diversity in Animals. Science 312, 570–572. doi:10.1126/science.1122033

Bearhop, S., Fiedler, W., Furness, R.W., Votier, S.C., Waldron, S., Newton, J., Bowen, G.J., Berthold, P., Farnsworth, K., 2005. Assortative Mating as a Mechanism for Rapid Evolution of a Migratory Divide. Science 310, 502–504. doi:10.1126/science.1115661

Betrán, E., Rozas, J., Navarro, A., Barbadilla, A., 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146, 89–99.

Bilton, D.T., Mirol, P.M., Mascheretti, S., Fredga, K., Zima, J., Searle, J.B., 1998. Mediterranean Europe as an area of endemism for small mammals rather than a source for northwards postglacial colonization. Proceedings of the Royal Society of London. Series B: Biological Sciences 265, 1219–1226.

Bradshaw, W.E., Fujiyama, S., Holzapfel, C.M., 2000. Adaptation to the thermal climate of North America by the pitcher-plant mosquito, Wyeomyia smithii. Ecology 81, 1262–1272.

Clark, P.U., Dyke, A.S., Shakun, J.D., Carlson, A.E., Clark, J., Wohlfarth, B., Mitrovica, J.X., Hostetler, S.W., McCabe, A.M., 2009. The last glacial maximum. science 325, 710–714.

Colangelo, P., Aloise, G., Franchini, P., Annesi, F., Amori, G., 2012. Mitochondrial DNA reveals hidden diversity and an ancestral lineage of the bank vole in the Italian peninsula. J Zool 287, 41–52. doi:10.1111/j.1469-7998.2011.00884.x

Cook, A., 1975. Changes in the Carrion/Hooded Crow hybrid zone and the possible importance of climate. Bird Study 22, 165–168.

Cronn, R., Cedroni, M., Haselkorn, T., Grover, C., Wendel, J.F., 2002. PCR-mediated recombination in amplification products derived from polyploid cotton. Theoretical and Applied Genetics 104, 482–489.

D'Amen, M., Zimmermann, N.E., Pearman, P.B., 2013. Conservation of phylogeographic lineages under climate change. Global Ecology and Biogeography 22, 93–104.

Dan, M., Hagiwara, A., 1967. Detection of two types of hemoglobin (HbA and HbF) in single erythrocytes by fluorescent antibody technique. Experimental Cell Research 46, 596–598. doi:10.1016/0014-4827(67)90385-0

Deffontaine, V., Ledevin, R., Fontaine, M.C., QUÉRÉ, J.-P., Renaud, S., Libois, R., Michaux, J.R., 2009. A relict bank vole lineage highlights the biogeographic history of the Pyrenean region in Europe. Molecular Ecology 18, 2489–2502.

Deffontaine, V., Libois, R., Kotlík, P., Sommer, R., Nieberding, C., Paradis, E., Searle, J.B., Michaux, J.R., 2005. Beyond the Mediterranean peninsulas: evidence of central European glacial refugia for a temperate forest mammal species, the bank vole (Clethrionomys glareolus). Molecular Ecology 14, 1727–1739. doi:10.1111/j.1365-294X.2005.02506.x

Fakruddin, M., Chowdhury, A., 2012. Pyrosequencing-An Alternative to Traditional Sanger Sequencing. American Journal of Biochemistry and Biotechnology 8, 14–20. doi:10.3844/ajbbsp.2012.14.20

Filipi, K., Marková, S., Searle, J.B., Kotlík, P., 2015. Mitogenomic phylogenetics of the bank vole Clethrionomys glareolus, a model system for studying end-glacial colonization of Europe. Molecular Phylogenetics and Evolution 82, Part A, 245–257. doi:10.1016/j.ympev.2014.10.016

Filipovska, A., Murphy, M.P., 2006. Overview of Protein Glutathionylation, in: Current Protocols in Toxicology. John Wiley & Sons, Inc.

Haldane, J.B.S., 1954. An exact test for randomness of mating. J Genet 52, 631–635. doi:10.1007/BF02985085

Hall, S.J.G., 1979. Haemoglobin polymorphism in the Bank vole, Clethrionomys glareolus, in Britain. Journal of Zoology 187, 153–160. doi:10.1111/j.1469-7998.1979.tb03939.x

Hartmann, D.L., Klein Tank, A.M.G., Rusticucc, M., Alexander, L.V., Brönnimann, S., Charabi, Y., Dentener, F.J., Dlugokencky, E.J., Easterling, D., Kaplan, A., Soden, B.J., Thorne, P.W., Wild, M., Zhai, P.M., 2013. Observations: Atmosphere and Surface., in: Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, and New York, NY, USA.

Hewitt, G., 2000. The genetic legacy of the Quaternary ice ages. Nature 405, 907–913.

Hewitt, G.M., 2004. Genetic consequences of climatic oscillations in the Quaternary. Philosophical Transactions of the Royal Society of London B: Biological Sciences 359, 183–195.

Hewitt, G.M., 1999. Post-glacial re-colonization of European biota. Biological Journal of the Linnean Society 68, 87–112.

Hofreiter, M., Münzel, S., Conard, N.J., Pollack, J., Slatkin, M., Weiss, G., Pääbo, S., 2007. Sudden replacement of cave bear mitochondrial DNA in the late Pleistocene. Current Biology 17, R122–R123.

Horáček, I., 2000. Glacial Cycles and Mammalian Biodiversity of Central Europe: Large Scale Migrations or Vicariance Dynamics. GeoLines 11, 103 – 107.

Jacob, M.H., Amir, D., Ratner, V., Gussakowsky, E., Haas, E., 2005. Predicting reactivities of protein surface cysteines as part of a strategy for selective multiple labeling. Biochemistry 44, 13664–13672.

Johns, G.C., Avise, J.C., 1998. A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene. Molecular Biology and Evolution 15, 1481–1490.

Jump, A.S., Hunt, J.M., Martínez-Izquierdo, J.A., Penuelas, J., 2006. Natural selection and climate change: temperature-linked spatial and temporal trends in gene frequency in Fagus sylvatica. Molecular Ecology 15, 3469–3480.

Keller, I., Taverna, A., Seehausen, O., 2011. A pyrosequencing assay for the rapid discrimination of mitochondrial lineages in the Salmo trutta species complex. Molecular ecology resources 11, 196–199.

Kellermann, V., Loeschcke, V., Hoffmann, A.A., Kristensen, T.N., Fløjgaard, C., David, J.R., Svenning, J.-C., Overgaard, J., 2012. Phylogenetic Constraints in Key Functional Traits Behind Species' Climate Niches: Patterns of Desiccation and Cold Resistance Across 95 Drosophila Species. Evolution 66, 3377–3389. doi:10.1111/j.1558-5646.2012.01685.x

Kelly, M.W., Sanford, E., Grosberg, R.K., 2012. Limited potential for adaptation to climate change in a broadly distributed marine crustacean. Proceedings of the Royal Society of London B: Biological Sciences 279, 349–356. doi:10.1098/rspb.2011.0542

Kotlík, P., Deffontaine, V., Mascheretti, S., Zima, J., Michaux, J.R., Searle, J.B., 2006. A northern glacial refugium for bank voles (Clethrionomys glareolus). PNAS 103, 14860–14864. doi:10.1073/pnas.0603237103

Kotlík, P., Marková, S., Vojtek, L., Stratil, A., Šlechta, V., Hyršl, P., Searle, J.B., 2014. Adaptive phylogeography: functional divergence between haemoglobins derived from different glacial refugia in the bank vole. Proc. R. Soc. B 281, 20140021. doi:10.1098/rspb.2014.0021

Kvist, L., Martens, J., Nazarenko, A.A., Orell, M., 2003. Paternal leakage of mitochondrial DNA in the great tit (Parus major). Mol. Biol. Evol. 20, 243–247.

Lavington, E., Cogni, R., Kuczynski, C., Koury, S., Behrman, E.L., O'Brien, K.R., Schmidt, P.S., Eanes, W.F., 2014. A Small System—High-Resolution Study of Metabolic Adaptation in the Central Metabolic Pathway to Temperate Climates in Drosophila melanogaster. Mol Biol Evol 31, 2032–2041. doi:10.1093/molbev/msu146

Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451–1452. doi:10.1093/bioinformatics/btp187

Macdonald, D.W., Barrett, P., 1993. Mammals of Britain & Europe. HarperCollins.

Magri, D., Vendramin, G.G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., Latałowa, M., Litt, T., Paule, L., Roure, J.M., Tantau, I., Van Der Knaap, W.O., Petit, R.J., De Beaulieu, J.-L., 2006. A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. New Phytologist 171, 199–221. doi:10.1111/j.1469-8137.2006.01740.x

Marková, S., Searle, J.B., Kotlík, P., 2014. Relaxed functional constraints on triplicate α-globin gene in the bank vole suggest a different evolutionary history from other rodents. Heredity 113, 64–73. doi:10.1038/hdy.2014.12

Miranda, J.J., 2000. Highly reactive cysteine residues in rodent hemoglobins. Biochemical and biophysical research communications 275, 517–523.

Mozzarelli, A., Bruno, S., Ronda, L., 2013. Biochemistry of Hemoglobin, in: Kim, H.W., Greenburg, A.G. (Eds.), Hemoglobin-Based Oxygen Carriers as Red Cell Substitutes and Oxygen Therapeutics. Springer Berlin Heidelberg, pp. 55–73.

Nussey, D.H., Postma, E., Gienapp, P., Visser, M.E., 2005. Selection on heritable phenotypic plasticity in a wild bird population. Science 310, 304–306. doi:10.1126/science.1117004

Overgaard, J., Kearney, M.R., Hoffmann, A.A., 2014. Sensitivity to thermal extremes in Australian Drosophila implies similar impacts of climate change on the distribution of widespread and tropical species. Glob Change Biol 20, 1738–1750. doi:10.1111/gcb.12521

Parmesan, C., Yohe, G., 2003. A globally coherent fingerprint of climate change impacts across natural systems. Nature 421, 37–42. doi:10.1038/nature01286

Piccinini, M., Kleinschmidt, T., Gorr, T., Weber, R.E., Künzle, H., Braunitzer, G., 1991. Primary structure and oxygen-binding properties of the hemoglobin from the lesser hedgehog tenrec (Echinops telfairi, Zalambdodonta). Evidence for phylogenetic isolation. Biological chemistry Hoppe-Seyler 372, 975–990.

Piertney, S.B., Stewart, W.A., Lambin, X., Telfer, S., Aars, J., Dallas, J.F., 2005. Phylogeographic structure and postglacial evolutionary history of water voles (Arvicola terrestris) in the United Kingdom. Molecular ecology 14, 1435–1444.

Pond, S.L.K., Frost, S.D., 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21, 2531–2533.

Pond, S.L.K., Muse, S.V., 2005. HyPhy: Hypothesis Testing Using Phylogenies, in: Statistical Methods in Molecular Evolution, Statistics for Biology and Health. Springer New York, pp. 125–181.

Pond, S.L.K., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006a. GARD: a genetic algorithm for recombination detection. Bioinformatics 22, 3096–3098.

Pond, S.L.K., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006b. Automated phylogenetic detection of recombination using a genetic algorithm. Molecular biology and evolution 23, 1891–1901.

Pulido, F., Berthold, P., Mohr, G., Querner, U., 2001. Heritability of the timing of autumn migration in a natural bird population. Proc Biol Sci 268, 953–959. doi:10.1098/rspb.2001.1602

QIAGEN, 2010. Pyrosequencing-The synergy of sequencing and quantification.

Rane, R.V., Rako, L., Kapun, M., Lee, S.F., Hoffmann, A.A., 2015. Genomic evidence for role of inversion 3RP of Drosophila melanogaster in facilitating climate change adaptation. Mol Ecol n/a–n/a. doi:10.1111/mec.13161

Réale, D., McAdam, A.G., Boutin, S., Berteaux, D., 2003. Genetic and plastic responses of a northern mammal to climate change. Proc Biol Sci 270, 591–596. doi:10.1098/rspb.2002.2224

Reischl, E., Dafre, A.L., Franco, J.L., Wilhelm Filho, D., 2007. Distribution, adaptation and physiological meaning of thiols from vertebrate hemoglobins. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 146, 22–53. doi:10.1016/j.cbpc.2006.07.015

Rohling, E.J., Fenton, M., Jorissen, F.J., Bertrand, P., Ganssen, G., Caulet, J.P., 1998. Magnitudes of sea-level lowstands of the past 500,000 years. Nature 394, 162–165.

Ronaghi, M., 2001. Pyrosequencing Sheds Light on DNA Sequencing. Genome Res. 11, 3–11. doi:10.1101/gr.150601

Root, T.L., Price, J.T., Hall, K.R., Schneider, S.H., Rosenzweig, C., Pounds, J.A., 2003. Fingerprints of global warming on wild animals and plants. Nature 421, 57–60. doi:10.1038/nature01333

Rossi, R., Barra, D., Bellelli, A., Boumis, G., Canofeni, S., Di Simplicio, P., Lusini, L., Pascarella, S., Amiconi, G., 1998. Fast-reacting thiols in rat hemoglobins can intercept damaging species in erythrocytes more efficiently than glutathione. Journal of Biological Chemistry 273, 19198–19206.

Rousset, F., 2008. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources 8, 103–106. doi:10.1111/j.1471-8286.2007.01931.x

Rousset, F., Raymond, M., 1995. Testing heterozygote excess and deficiency. Genetics 140, 1413–1419.

Royo, J.L., Hidalgo, M., Ruiz, A., 2007. Pyrosequencing protocol using a universal biotinylated primer for mutation detection and SNP genotyping. Nature Protocols 2, 1734–1739. doi:10.1038/nprot.2007.244

Rychlik, W., 2007. OLIGO 7 Primer Analysis Software, in: Yuryev, A. (Ed.), PCR Primer Design, Methods in Molecular Biology™. Humana Press, pp. 35–59.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74, 5463–5467.

Searle, J.B., Kotlík, P., Rambau, R.V., Marková, S., Herman, J.S., McDevitt, A.D., 2009. The Celtic fringe of Britain: insights from small mammal phylogeography. Proceedings of the Royal Society of London B: Biological Sciences rspb20091422. doi:10.1098/rspb.2009.1422

Searle, J.B., Wilkinson, P.J., 1987. Karyotypic variation in the common shrew (Sorex araneus) in Britain—a "Celtic fringe." Heredity 59, 345–351.

Sha, J., Liang, G., Pan, J., Xuan, H., Ping, P., Li, D., Bo, J., Liu, D., Shen, W., Liu, W., Huang, Y., 2012. Application of pyrosequencing technique for improved detection of K-Ras mutation in formalin-fixed and paraffin-embedded prostate carcinoma tissues in Chinese patients. Clinica Chimica Acta 413, 1532–1535. doi:10.1016/j.cca.2012.06.008

Soares, P., Abrantes, D., Rito, T., Thomson, N., Radivojac, P., Li, B., Macaulay, V., Samuels, D.C., Pereira, L., 2013. Evaluating Purifying Selection in the Mitochondrial DNA of Various Mammalian Species. PLoS ONE 8, e58993. doi:10.1371/journal.pone.0058993

Spinney, L., 2008. Archaeology: The lost world. Nature News 454, 151–153. doi:10.1038/454151a

Stephens, M., Donnelly, P., 2003. A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. Am J Hum Genet 73, 1162–1169.

Stephens, M., Smith, N.J., Donnelly, P., 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. Am J Hum Genet 68, 978–989.

Stewart, J.R., Lister, A.M., 2001. Cryptic northern refugia and the origins of the modern biota. Trends in Ecology & Evolution 16, 608–613.

Stewart, J.R., Lister, A.M., Barnes, I., Dalén, L., 2010. Refugia revisited: individualistic responses of species in space and time. Proceedings of the Royal Society B: Biological Sciences 277, 661–671.

Storz, J.F., Runck, A.M., Sabatino, S.J., Kelly, J.K., Ferrand, N., Moriyama, H., Weber, R.E., Fago, A., 2009. Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin. Proceedings of the National Academy of Sciences 106, 14450–14455.

Storz, J.F., Sabatino, S.J., Hoffmann, F.G., Gering, E.J., Moriyama, H., Ferrand, N., Monteiro, B., Nachman, M.W., 2007. The molecular basis of high-altitude adaptation in deer mice. PLoS Genetics 3, e45.

Sundquist, A., Bigdeli, S., Jalili, R., Druzin, M.L., Waller, S., Pullen, K.M., El-Sayed, Y.Y., Taslimi, M.M., Batzoglou, S., Ronaghi, M., 2007. Bacterial flora-typing with targeted, chip-based Pyrosequencing. BMC Microbiology 7, 108. doi:10.1186/1471-2180-7-108

Taberlet, P., FUMAGALLI, L., WUST-SAUCY, A.-G., COSSON, J.-F., 1998. Comparative phylogeography and postglacial colonization routes in Europe. Molecular Ecology 7, 453–464.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729. doi:10.1093/molbev/mst197

Tatarenkov, A., Avise, J.C., 2007. Rapid concerted evolution in animal mitochondrial DNA. Proceedings of the Royal Society of London B: Biological Sciences 274, 1795–1798. doi:10.1098/rspb.2007.0169

Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T.J., Fromentin, J.-M., Hoegh-Guldberg, O., Bairlein, F., 2002. Ecological responses to recent climate change. Nature 416, 389–395.

Weber, R.E., Fago, A., 2004. Functional adaptation and its molecular basis in vertebrate hemoglobins, neuroglobins and cytoglobins. Respiratory Physiology & Neurobiology, Frontiers in Comparative Respiratory Physiology 1: Adaptations in Respiratory Gas Transport 144, 141–159. doi:10.1016/j.resp.2004.04.018

Weninger, B., Schulting, R., Bradtmöller, M., Clare, L., Collard, M., Edinborough, K., Hilpert, J., Jöris, O., Niekus, M., Rohling, E.J., others, 2008. The catastrophic final flooding of Doggerland by the Storegga Slide tsunami. Documenta Praehistorica 35, 1–24.

Wójcik, J.M., Kawałko, A., Marková, S., Searle, J.B., Kotlík, P., 2010. Phylogeographic signatures of northward post-glacial colonization from high-latitude refugia: a case study of bank voles using museum specimens. Journal of Zoology 281, 249–262. doi:10.1111/j.1469-7998.2010.00699.x

Annexes

I Annexe 1: List of regional populations and corresponding coordinates of regional populations centres, number of samples included in the populations and HBB-T1 and HBB-T2 52Cys allele frequency in each population.

II Annexe 2: Optimized PCR protocols for pyrosequencing analysis with final selected set of amplification primers for both the HBB-T1 and HBB-T2 genes.

III Annexe 3a: Protocol for PCR product immobilization, DNA pyrosequencing template purification and annealing of sequencing primers provided by Qiagen, 1st page.

IV Annexe 3b: Protocol for PCR product immobilization, DNA pyrosequencing template purification and annealing of sequencing primers provided by Qiagen, 2nd page.

V Annexe 4a: Optimized reagents mixes for immobilization of PCR products and annealing of sequencing primers.

Annexe 4b: Example of pyrosequencing run preparation information generated prior to each run by PyroMark software. Reagents amounts and scheme of loading into the dispensation cartridge is shown.

VI Annexe 5: List of samples selected for Sanger sequencing.

VII Annexe 6a: Optimized PCR protocol for Sanger sequencing analysis.

Annexe 6b: Sequences of sequencing primers used for Sanger sequencing.

VIII Annexe 7a: List of primers used for correction runs.

Annexe 7b: List of samples exhibiting missing peaks.

IX Annexe 8: Complete list of HBB-T1 and HBB-T2 haplotypes, samples included in each haplotype and sampling locality for each sample.

X Annexe 9: Result of gene conversion analyses in the form of reduced table.

XI Annexe 10: Poster on which part of my results was presented at Rodens et Spatium conference in Lisbon in 2014.

XII Annexe 11. Article of P. Kotlík about bank vole haemoglobin polymorphism in Czech popular scientific magazine Živa for which I am part of the co-author team.

Annexe 1: List of regional populations and corresponding coordinates of regional populations centres, number of samples included in the populations and HBB-T1 and HBB-T2 52Cys allele frequency in each population.

				52Cys frquency	
Regional population	Latitude	Longitude	nr. of samples	HBB-T1	HBB-T2
AT1	47.85	15.05	10	0.000	0.000
BEL1	50.63	4.18	5	1.000	1.000
BGR1	43.10	23.40	4	0.250	0.125
BGR2	42.20	23.00	5	0.600	0.100
BGR3	41.50	24.50	3	0.833	0.167
BGR4	42.913	27.0533	3	0.833	0.333
CZAT	48.84	16.66	11	0.818	0.000
CZE1	50.22909	13.95182	11	0.909	0.591
CZE2	49.565	14.90417	12	0.417	0.125
CZE3	50.50	16.12	6	0.417	0.000
CZE4	49.75	18.30	10	0.100	0.100
CZSK	48.86	18.58	11	0.545	0.000
DE1	53.62	7.27	10	1.000	0.200
DE2	54.22	9.11	10	0.150	0.000
DE3	53.65043	10.783578	12	0.917	0.417
DE4	51.852	11.181	10	1.000	0.950
DE5	51.86	13.06	10	1.000	1.000
DE6	50.7111	11.8651	3	0.667	0.500
DE7	49.02	10.73	7	1.000	0.143
DNK1	55.7968	9.4076	5	0.000	0.000
DNK2	54.96376	11.62884	7	0.000	0.000
FRA1	47.34	-3.17	1	-	1.000
FRA2	48.66	-1.50	4	1.000	1.000
FRA3	48.97	0.13	5	1.000	1.000
FRA4	50.576	1.801	8	1.000	0.938
FRA5	48.70625	1.7625	8	1.000	1.000
FRA6	46.04364	2.59545	11	0.955	0.455
FRA7	43.264762	0.249286	13	0.731	0.071
HR1	45.74	15.65	6	0.000	0.000
HR2	44.839	15.62	10	0.000	0.000
HU1	47.30	19.22	7	0.000	0.000
CHE1	46.40	6.3763	8	0.062	0.000
CHE2	45.90	7.20	6	0.000	0.000
CHE3	47.37	8.55	2	0.750	0.000
CHEATIT	46.78154	10.26769	13	0.077	0.000
IRL1	53.11	-8.60	2	1.000	0.500
IT1	42.8668	11.5936	6	0.000	0.000
IT2	41.81692	16.00904	4	0.000	0.000
IT3	39.34744	16.45366	10	0.000	0.050
NLD1	52.500727	4.573636	11	0.955	0.773
NLD2	53.072	6.244	5	1.000	0.100
NLD3	52.071	6.076	5	1.000	0.200
NOR1	66.42091	14.7705	1	0.000	0.000
NOR2	64.14931	11.97037	6	0.000	0.000
NOR3	62.96309	8.71408	1	0.000	0.000
NOR4	60.79039	5.97426	8	0.000	0.000
NOR5	61.43184	10.13544	8	0.000	0.000
NOR6	58.34787	8.26175	11	0.000	0.000
PL1	52.45	17.10	4	0.500	0.375
PL2	51.47	19.82	4	0.750	0.500
PL3	53.80	21.65	10	0.000	0.000
PL4	52.73	23.85	10	0.000	0.000
RO1	46.483	23.655	6	0.333	0.000
RO2	46.55	26.83	6	0.000	0.000
RUS1	57.15	33.10	5	0.000	0.000
SK1	48.38364	19.08091	11	0.682	0.045
SK2	49.1591	20.0682	11	0.273	0.000
SK3	48.77	21.384	9	0.125	0.000
SRB1	44.09	21.64	4	0.625	0.125
SRB2	43.24	20.82	5	0.400	0.000
SRBHR	45.166	19.58	7	0.571	0.000
SVN1	46.11	14.67	6	0.000	0.000
SWE1	62.87948	17.73554	11	0.000	0.000
SWE2	60.16797	12.77189	10	0.200	0.050
SWE3	61.20140	16.585	11	0.409	0.000
SWE4	57.62071	12.41887	1	0.000	0.000
SWE5	59.20503	16.63308	10	0.700	0.050
SWE6	55.69928	13.47494	10	0.000	0.000
TUR1	40.12	29.17	10	0.000	0.000
UA1	50.4083	29.7167	6	0.000	0.000

Annexe 2: Optimized PCR protocols for pyrosequencing analysis with final selected set of amplification primers for both the HBB-T1 and HBB-T2 genes.

PCR N --
Bankvoles- Beta Globin paralog T1

HBB_T1_RB_F19	5' GGCTGCTGGTKGTCTACCC 3'
HBB_T1_RB_R19	5' ACAGGCAAGTGCAGGAAAG 3'

Program **T160LONG**

1=94.0°	for	2:00
2=94.0°	for	0:40
3=**60.0**°	for	1:00
4=72.0°	for	1:30
5=Goto 2, 33 times		
6=72.0°	for	10:00
7=4.0°	for	ever

PPP Master Mix

vz.	1	X
PPP Master Mix	20	20x
L primer	2	2x
H primer	2	2x
H2O	13	13x
	37	
DNA	3	
Total	40	40x

PCR N --
Bankvoles- Beta Globin paralog T2

HBB_T2_RB_F21	5' TTGGACCCAGAGGTTCTTTGA 3'
HBB_T2_RB_R21	5' CTCCAGGGGACAAAAAACATT 3'

Program **T262**

1=94.0°	for	2:00
2=94.0°	for	0:40
3=**62.0**°	for	1:00
4=72.0°	for	1:30
5=Goto 2, 33 times		
6=72.0°	for	10:00
7=4.0°	for	ever

PPP Master Mix

vz.	1	X
PPP Master Mix	20	20x
L primer	2	2x
H primer	2	2x
H2O	13	13x
	37	
DNA	3	
Total	40	40x

Annexe 3a: Protocol for PCR product immobilization, DNA pyrosequencing template purification and annealing of sequencing primers provided by Qiagen, 1st page.

TechnicalInformation

PyroMark® Q24 Vacuum Workstation quick-start guide

This Technical Information summarizes the immobilization and preparation of PCR products for Pyrosequencing® on the PyroMark Q24 Instrument. Before using the PyroMark Q24 Vacuum Workstation, carefully read Section 5.3 of the *PyroMark Q24 User Manual* and pay particular attention to the safety information.

Immobilizing the PCR products

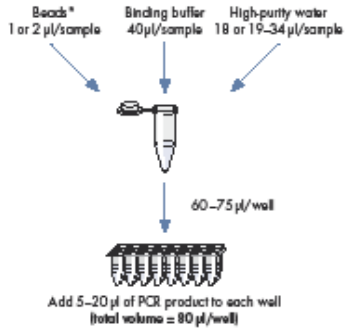
1. Make a master mix according to the flowchart to the right.

   Note: Before pipetting, gently shake the bottle of streptavidin-coated Sepharose® beads* to ensure a homogenous suspension.

2. Depending on the sample volume, dispense 60–75 µl master mix into each necessary well of a PCR plate to give a total volume of 80 µl per well.

3. Add 5–20 µl PCR product to each well.

4. Seal the wells with strip caps and agitate the PCR plate at 1400 rpm for 5–10 min at room temperature (15–25°C) using an orbital shaker.



Preparing the master mix to immobilize the PCR products.

Separation of DNA strands and release of samples into the PyroMark Q24 Plate

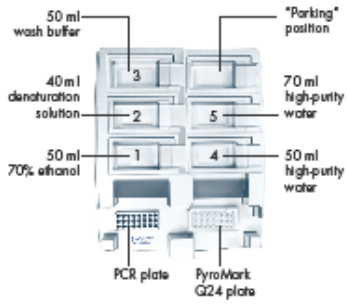
1. Dilute the sequencing primers to 0.3 µM with PyroMark Annealing Buffer (cat. no. 979009), and dispense 25 µl into each necessary well of a PyroMark Q24 Plate. Position the plate on the workstation.

2. Fill the workstation troughs according to the diagram to the right.

3. Start the pump and apply vacuum to the tool by opening the switch. Flush the filter probes with high-purity water (Milli-Q® 18.2 MΩ x cm or equivalent) in trough 5. Refill the trough with fresh high-purity water for use in step 12.

4. Position the PCR plate on the workstation. Ensure that both plates are in the same orientation as when the samples were loaded.

5. With the vacuum switch ON, lower the vacuum tool into the wells of the PCR plate for 15 s to capture the beads with PCR product.



Filling the workstation troughs.

* Streptavidin Sepharose High Performance (34 µm, 5 ml, GE Healthcare). Check the lot number of the Streptavidin Sepharose High Performance. For lot number 10057037 or higher, use 1 µl in the master mix. For lot numbers lower than 10057037, use 2 µl. This is not a complete list of suppliers and does not include many important vendors of biological supplies.

QIAGEN®

Sample & Assay Technologies

Annexe 3b: Protocol for PCR product immobilization, DNA pyrosequencing template purification and annealing of sequencing primers provided by Qiagen, 2nd page.

CAUTION:

PyroMark Denaturation Solution contains sodium hydroxide, which is an eye and skin irritant.

Always wear a suitable lab coat, disposable gloves, and protective goggles. For more information, see the *PyroMark Q24 User Manual* (and the MSDS).

Vacuum switch ON.

Vacuum switch OFF.

6. With the vacuum ON, flush the tool with 70% ethanol (trough 1) for 5 s.
7. With vacuum ON, flush the tool with denaturation solution (trough 2) for 5 s.
8. With vacuum ON, flush the tool with wash buffer (trough 3) for 10 s.
9. With vacuum ON, raise the vacuum tool to beyond 90° vertical for 5 s.
10. Align the vacuum tool with the PyroMark Q24 Plate and switch the vacuum OFF. Lower the vacuum tool into the wells and gently shake from side to side to release the beads.
11. With the vacuum OFF, agitate the vacuum tool in high-purity water (trough 4) for 10 s.
12. Switch the vacuum ON and flush the filter probes with high-purity water (trough 5) for 5 s.
13. Raise the vacuum tool to beyond 90° vertical for 5 s, then switch the vacuum OFF and store the tool in the "Parking" position.

Annealing sequencing primers to DNA strands

Place the PyroMark Q24 Plate in a prewarmed PyroMark Q24 Plate Holder. Heat the Pyrosequencing samples on a heating block at 80°C for 2 minutes.

Remove the plate from the holder and allow the samples to cool to room temperature (15–25°C) for at least 5 minutes. The cooled plate can now be processed in the PyroMark Q24.

Cleaning the vacuum workstation

Liquid waste and solutions remaining in the troughs of the vacuum workstation should be appropriately discarded at the end of the day. For details, see the *PyroMark Q24 User Manual*.

For up-to-date licensing information and product-specific disclaimers, see the respective *PyroMark Q24 User Manual*. The *PyroMark Q24 User Manual* is available at www.qiagen.com or can be requested from QIAGEN Technical Services or your local distributor.

Australia = 1-800-243-800
Austria = 0800-281011
Belgium = 0800-79612
Brazil = 0800-557779
Canada = 800-572-9613
China = 800-988-0325
Denmark = 80-885945
Finland = 0800-914416
France = 01-60-920-930
Germany = 02103-29-12000
Hong Kong = 800 933 965
India = 1-800-102-4114
Ireland = 1800 555 049
Italy = 800-787980
Japan = 03-6890-7300
Korea (South) = 080-000-7145
Luxembourg = 8002 2076
Mexico = 01-800-7742-436
The Netherlands = 0800-0229592
Norway = 800-18859
Singapore = 1800-742-4368
Spain = 91-630-7050
Sweden = 020-790282
Switzerland = 055-254-22-11
Taiwan = 0080-665-1947
UK = 0808-234-3665
USA = 800-426-8157
www.qiagen.com

Sample & Assay Technologies

Annexe 4a: Optimized reagents mixes for immobilization of PCR products and annealing of sequencing primers.

Immobilizing the PCR products

vz.	1	26
Binding buffer	40	1040
Beads	3	78
H2O	0	0
PCR product (10-20nM)	37	
Total	80	2080

Annealing buffer + seq.primer

vz.	1	26
annealing buffer	23.5	611
seq.primer	1.5	39
total	25	650

Immobilizing, purification and annealing procedures were followed as described in protocol in Annexe 3.

Annexe 4b: Example of pyrosequencing run preparation information generated prior to each run by PyroMark software. Reagents amounts and scheme of loading into the dispensation cartridge is shown.

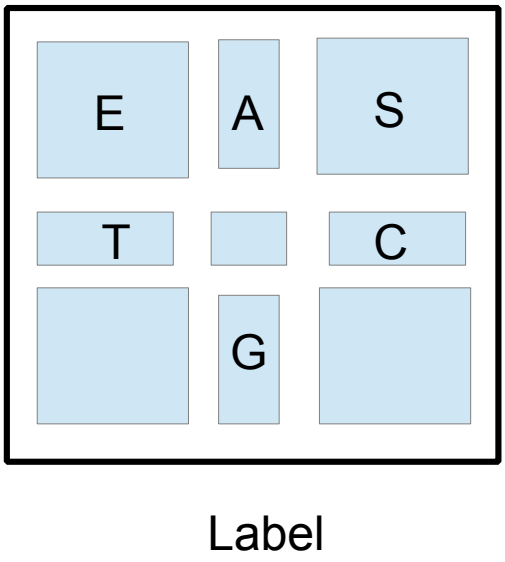
Volume Information:

Reagents:

E-mix: 102 µl
S-mix: 102 µl

Nucleotides:

Nucleotide A: 67 µl
Nucleotide C: 67 µl
Nucleotide G: 71 µl
Nucleotide T: 67 µl



Annexe 5: List of samples selected for Sanger sequencing.

Sample	Sampling locality	Country	HBB-T1	HBB-T2
18	Feofania Forest, Kiev	Ukraine	Serin	Serin
271	Solnik, Varna, Balkan Mts (197m)	Bulgaria	Cystein	Cystein/Serin
273	Elena, Veliko Tarnovo, Balkan Mts (618m)	Bulgaria	Cystein/Serin	Serin
274	Bacau (284m)	Romania	Serin	Serin
450	Strmosten, Kločanica River gorge, Resava River basin, Beljanica Mts (333m)	Serbia and Mo	Cystein/Serin	Cystein/Serin
498	Vranovice (166m)	Czech Rep	Cystein/Serin	Serin
601	Bezau, Vorarlberg (745m)	Austria	Cystein	Serin
622	Mšec, Central Bohemia (378m)	Czech Rep	Cystein	Cystein/Serin
623	Mšec, Central Bohemia (378m)	Czech Rep	Cystein/Serin	Serin
649	Bursa, Uludag Mts, Marmara Region, Western Anatolia	Turkey	Serin	Serin
707	Hakel forest, Quedlinburg, Sachsen-Anhalt, NE Harz foreland	Germany	Cystein	Cystein
734	Polichno, Piotrków Trybunalski, Central European Lowland (192m)	Poland	Cystein	Cystein/Serin
735	Polichno, Piotrków Trybunalski, Central European Lowland (192m)	Poland	Cystein/Serin	Serin
737	Bialowieza National Park	Poland	Serin	Serin
755	Ostashkhov, Seliger Lake	Russia	Serin	Serin
836	Beauvoir	France	Cystein	Cystein
1046	Geneva, Jura Mts, Alps	Switzerland	Cystein/Serin	
1062	Zurich, Alps	Switzerland	Cystein/Serin	Serin
1104	Ardennes	Belgium	Cystein	Cystein
1126	Bussiéres; La Cote	France	Cystein	Cystein/Serin
1187	Hanišberg, Krupina, Krupinska planina Mts, Slovak Medium Mts (350m)	Slovakia	Cystein	Cystein/Serin
1188	Hanišberg, Krupina, Krupinska planina Mts, Slovak Medium Mts (350m)	Slovakia	Cystein/Serin	Serin
1327	Ratzerburg, Moeln, Schleswig-Holstein	Germany	Cystein	Cystein/Serin
1408	St. Omer, Nord Pas de Calais	France	Cystein	Cystein/Serin
1433	Pyrenees	France	Serin	Serin
1435	Pyrenees	France	Cystein/Serin	Serin
1517	Noordwijk, Nieuw Leeuwenhorst, South Holland (-21m)	Netherlands	Cystein	Cystein/Serin
1641	Skog	Sweden	Cystein/Serin	Serin
1698	Arla 2	Sweden	Cystein	Cystein/Serin
1742	Torsby3, ve svahu u kamenu s borvkami	Sweden	Cystein/Serin	Cystein/Serin
1765	Strandkaer, Molslaboratoriet	Denmark	Serin	Serin
1826	Calabria, Sila Grande, Valle Capra	Italy	Serin	Serin
1828	Calabria, Sila Grande, Valle Capra	Italy	Serin	Cystein/Serin
1845	Apulia, Gargano peninsula, Foresta Umbra	Italy	Serin	Serin
1848	Snåsa, Nord-Trøndelag	Norway	Serin	Serin
1867	Eksingedal, Vaksdal, Eksingedalen, Hordaland	Norway	Serin	Serin
1903	Viterbo	Italy	Serin	Serin
1910	Armendarits	France	Cystein	Cystein/Serin
1048	Geneva, Jura Mts, Alps	Switzerland		
1395	Ribeaville	France		
1670	Kramfors1, les smrk u cesty	Sweden	Serin	
1859	Venabygd, Ringebu, Oppland	Norway	Serin	
1871	Eksingedal, Vaksdal, Eksingedalen, Hordaland	Norway		
1879	Lillesand, Aust-Agder	Norway	Serin	
1911	Armendarits	France		Serin

Samples with sequences already available from previous studies are highlighted in green.

Samples discarded due to poor PCR amplification results are highlighted in orange.

Samples added to the list due to missing peaks in pyrosequencing analysis are marked by the black rectangle.

Annexe 6a: Optimized PCR protocol for Sanger sequencing analysis.

PCR N --
Bankvoles- beta Globin paralog T1

BT1F1 5‘ ACAYTTGCTTCTGACATAGT 3‘
BT1R593 5‘ TGAAAGTAAATGCCTTTTATTAGT 3‘

Program BETA T1

1=94.0°	for	2:00
2=94.0°	for	0:30
3=57.0°	for	0:40
4=72.0°	for	1:00
5=Goto 2, 33 times		
6=72.0°	for	10:00
7=12.0°	for	10:00

PPP Master Mix

vz.	1	26
PPP Master Mix	12,5	325
L primer	0,75	19,50
H primer	0,75	19,50
H2O	9,5	247,0
	23,5	
DNA	1,5	
Total	**25**	650

PCR N --
Bankvoles -beta globin paralog T2

HBB10U19 5‘ ATGCACACCCTGGAATTGG 3‘
HBB1266L21 5’ GTGCATAAACACGAGCAAGAA 3‘

Program **BETA T2**

1=94.0°	for	2:00
2=94.0°	for	0:30
3=57.0°	for	0:40
4=72.0°	for	1:00
5=Goto 2, 34 times		
6=72.0°	for	10:00
7=12 °	for	10:00

PPP Master Mix

vz.	1	9
PPP Master Mix	12,5	112,5
L primer	0,75	6,75
H primer	0,75	6,75
H2O	9,5	85,50
	23,5	
DNA	1,5	
Total	**25**	225

Annexe 6b: Sequences of sequencing primers used for Sanger sequencing.

HBB-T1: sequencing primers = amplification primers

HBB-T2: HBB10U19 5'AGTCACACCCTGGAATTGG3'

HBB597U19 5'CCCGTGGTTTCCTTCCTCT3'

Annexe 7a: List of primers used for correction runs.

primer	sequence	type	direction
HBB_T1_RB_17F20	5'CCTTGGACCCAGAGGTTCTT3'	amplification	F
HBB_T1_RB_1010R25	5'TGAAAGTAAATGCCTTTTATTAGT3'	amplification	R
HBB_T1_S47F21	5'CACTTTGGGGACCTGTCCTCT3'	sequencing	F
HBB_T2_RB_FS15	5'GGGACCTGTCCTCTG3'	sequencing	F

Annexe 7b: List of samples exhibiting missing peaks.

Sample	Sampling locality	Country
1048	Geneva, Jura Mts, Alps	Switzerland
1395	Ribeaville	France
1670	Kramfors1, les smrk u cesty	Sweden
1859	Venabygd, Ringebu, Oppland	Norway
1871	Eksingedal, Vaksdal, Eksingedalen, Hordal	Norway
1879	Lillesand, Aust-Agder	Norway
1911	Armendarits	France

haplotype	sample		country	locality
Hap 1	CG1_18_1	CG1_18_2	UKR	Feofania
Hap 2	CG1_271_1	CG1_271_2	BGR	Solnik
Hap 3	CG1_274_1		ROM	Bacau
Hap 4		CG1_274_2	ROM	Bacau
Hap 5	CG1_450_1		SRB	Strmosten
Hap 6		CG1_450_2	SRB	Strmosten
Hap 7	CG1_622_1		CZE	Mšec
		CG1_623_2	CZE	Mšec
	CG1_1126_1		FRA	Bussiéres
Hap 8		CG1_622_2	CZE	Mšec
Hap 9	CG1_623_1		CZE	Mšec
Hap 10	CG1_707_1	CG1_707_2	DEU	Hakel forest
		CG1_1126_2	FRA	Bussiéres
	CG1_1327_1	CG1_1327_2	DEU	Ratzerburg
	CG1_1517_1	CG1_1517_2	NLD	Noordwijk
Hap 11	CG1_734_1		POL	Polichno
		CG1_1188_2	SVK	Hanišberg
Hap 12		CG1_734_2	POL	Polichno
	CG1_1187_1	CG1_1187_2	SVK	Hanišberg
Hap 13	CG1_735_1		POL	Polichno
	CG1_1188_1		SVK	Hanišberg
	CG1_1433_1	CG1_1433_2	FRA	Pyrenees
Hap 14		CG1_735_2	POL	Polichno
Hap 15	CG1_737_1		POL	Bialowieza
Hap 16		CG1_737_2	POL	Bialowieza
		CG1_755_2	RUS	Ostashkov
Hap 17	CG1_755_1		RUS	Ostashkov
Hap 18	CG1_836_1	CG1_836_2	FRA	Beauvoir
Hap 19	CG1_1435_1		FRA	Pyrenees
Hap 20		CG1_1435_2	FRA	Pyrenees
Hap 21	CG1_1641_1		SWE	Skog
	CG1_1742_1		SWE	Torsby
Hap 22		CG1_1641_2	SWE	Skog
	CG1_1698_1	CG1_1698_2	SWE	Arla
		CG1_1742_2	SWE	Torsby
Hap 23	CG1_1670_1		SWE	Kramfors
Hap 24		CG1_1670_2	SWE	Kramfors
	CG1_1859_1	CG1_1859_2	NOR	Venebygd
	CG1_1871_1	CG1_1871_2	NOR	Eksingedal
	CG1_1879_1	CG1_1879_2	NOR	Lillesand
Hap 25	CG1_1765_1	CG1_1765_2	DNK	Strandkaer
		CG1_1848_2	NOR	Snasa
	CG1_1867_1	CG1_1867_2	NOR	Eksingedal
Hap 26	CG1_1826_1			
		CG1_1828_2	ITA	Calabria
Hap 27		CG1_1826_2		
Hap 28	CG1_1828_1		ITA	Calabria
Hap 29	CG1_1845_1	CG1_1845_2	ITA	Gargano
Hap 30	CG1_1848_1		NOR	Snasa
Hap 31 – C.rufocanus	CG1_1855_1			
Hap 32 – C.rufocanus		CG1_1855_2		
Hap 33	CG1_1903_1		ITA	Viterbo
Hap 34		CG1_1903_2	ITA	Viterbo
Hap 35	CG1_1048_1		CHE	Geneva
Hap 36		CG1_1048_2	CHE	Geneva
Hap 37	CG1_1395_1	CG1_1395_2	FRA	Ribeavile
Hap 38	CG1_1911_1	CG1_1911_2	FRA	Armendarits
Hap 39	CG2_273_1	CG2_273_2	BGR	Elena
Hap 40	CG2_274_1		ROM	Bacau
	CG2_601_1		AUT	Bezau
	CG2_1126_1		FRA	Bussiéres
	CG2_1435_1		FRA	Pyrenees
	CG2_1903_1		ITA	Viterbo
Hap 41		CG2_274_2	ROM	Bacau
		CG2_735_2	POL	Polichno
Hap 42	CG2_450_1		SRB	Strmosten
	CG2_1641_1		SWE	Skog
	CG2_1698_1		SWE	Arla
	CG2_1845_1	CG2_1845_2	ITA	Gargano
Hap 43	CG2_450_2		SER	Strmosten
Hap 44		CG2_601_2	AUT	Bezau
	CG2_735_1		POL	Polichno
	CG2_1433_1	CG2_1433_2	FRA	Pyrenees
	CG2_1517_1		NLD	Noordwijk
Hap 45	CG2_622_1		CZE	Mšec
Hap 46		CG2_622_2	CZE	Mšec
	CG2_707_1	CG2_707_2	DEU	Hakel forest
		CG2_1517_2	NLD	Noordwijk
Hap 47	CG2_623_1		CZE	Mšec
Hap 48		CG2_623_2	CZE	Mšec
Hap 49	CG2_734_1		POL	Polichno
Hap 50		CG2_734_2	POL	Polichno
Hap 51	CG2_737_1	CG2_737_2	POL	Bialowieza
Hap 52	CG2_836_1		FRA	Beauvoir
Hap 53		CG2_836_2	FRA	Beauvoir
Hap 54		CG2_1126_2	FRA	Bussiéres
Hap 55	CG2_1187_1		SVK	Hanišberg
Hap 56		CG2_1187_2	SVK	Hanišberg
		CG2_1698_2	SWE	Arla
		CG2_1742_2	SWE	Torsby
Hap 57	CG2_1188_1		SVK	Hanišberg
Hap 58		CG2_1188_2	SVK	Hanišberg
Hap 59	CG2_1327_1		DEU	Ratzerburg
Hap 60		CG2_1327_2	DEU	Ratzerburg
Hap 61		CG2_1435_2	FRA	Pyrenees
Hap 62		CG2_1641_2	SWE	Skog
Hap 63	CG2_1742_1		SWE	Torsby
Hap 64	CG2_1765_1	CG2_1765_2	DNK	Strandkaer
Hap 65	CG2_1826_1			
Hap 66		CG2_1826_2		
		CG2_1903_2	ITA	Viterbo
Hap 67	CG2_1828_1	CG2_1828_2	ITA	Calabria
Hap 68 – C.rufocanus	CG2_1855_1	CG2_1855_2		
Hap 69	CG2_1848_1		NOR	Snasa
	CG2_1867_1		NOR	Eksingedal
Hap 70		CG2_1848_2	NOR	Snasa
Hap 71		CG2_1867_2	NOR	Eksingedal
Hap 72	CG2_755_1		RUS	Ostashkov
Hap 73		CG2_755_2	RUS	Ostashkov
Hap 74	CG2_649_1		TUR	Bursa
	CG2_271_1		BGR	Solnik
Hap 75		CG2_649_2	TUR	Bursa
Hap 76		CG2_271_2	BGR	Solnik
Hap 77	CG2_1871_1		NOR	Eksingedal
Hap 78		CG2_1871_2	NOR	Eksingedal
		CG2_1670_2	SWE	Kramfors
		CG2_1879_2	NOR	Lillesand
Hap 79	CG2_1670_1		SWE	Kramfors
	CG2_1879_1		NOR	Lillesand
Hap 80	CG2_1859_1		NOR	Venebygd
Hap 81	CG2_1859_2		NOR	Venebygd

Annexe 8: Complete list of HBB-T1 and HBB-T2 haplotypes, samples included in each haplotype and sampling locality for each sample.

Sample labels should be read as follows:

e.g. CG1_18_1

CG1 = ***C****lethrionomys* ***g****lareolus*, HBB-T**1** gene

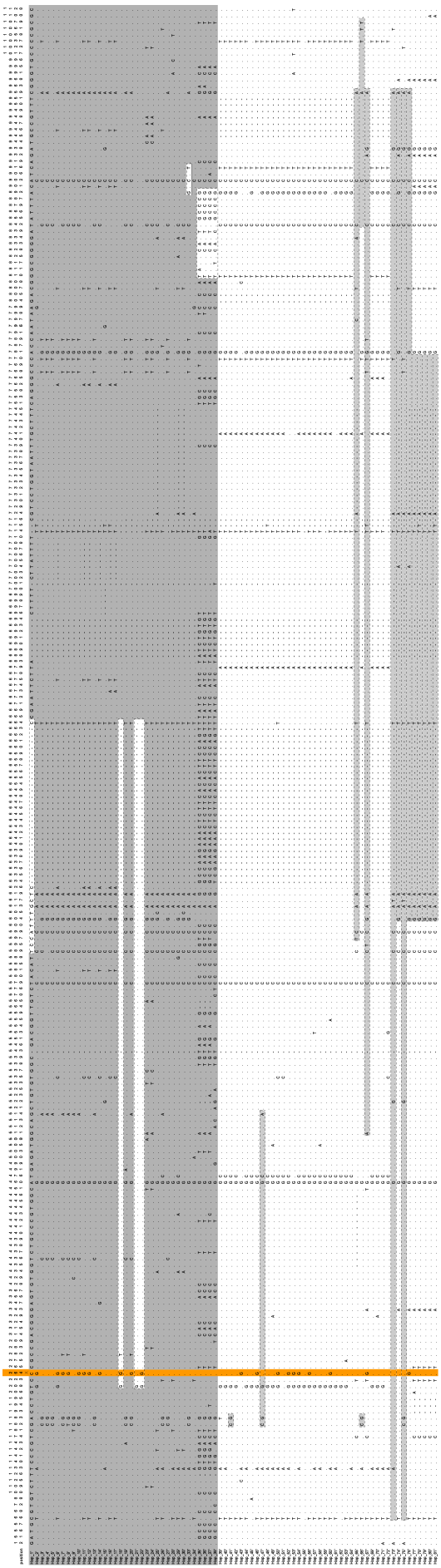
18 = sample label in stock

1 = allele 1

Alleles are recorded in two separate columns for easier orientation.

Hap 1 – Hap 38 belong to HBB-T1 gene
Hap 39 – Hap 81 belong to HBB-T2 gene

Hap 31, Hap 32 and Hap 68 are haplotypes of C. Rufocanus used as outgroup in phylogenetic analyses



Annexe 9: Result of gene conversion analyses in the form of reduced table.

For complete table see file Conversion.xls available on CD

Annexe 10: Poster on which part of my results was presented at Rodens et Spatium conference in Lisbon in 2014.

European origin of functional haemoglobin mutation related to population replacement in British bank voles

Michaela Strážnická[a,b], Silvia Marková[a], Jeremy B. Searle[c] and Petr Kotlík[a]

[a]Institute of Animal Physiology and Genetics AS CR, Liběchov, Czech Republic; kotlik@iapg.cas.cz
[b]Department of Animal Science and Food Processing, Czech University of Life Sciences Prague, Czech Republic
[c]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA

Functional haemoglobin polymorphism in the bank vole

Bank voles (*Clethrionomys glareolus*) can possess one of two genetically determined haemoglobin (Hb) variants differing by an amino acid substitution of serine (Ser) with cysteine (Cys) at position 52 in the beta haemoglobin chain (Kotlík et al. 2014). An exposed side chain sulfur atom makes Cys in Hb a highly reactive functional residue with an important role in red blood cell (RBC) protection against oxidative damage. This is likely because such Cys can take part in the regeneration of glutathione (GSH), a major intracellular antioxidant, and/or can intercept reactive oxygen species directly (Rossi et al. 1998). In the bank vole, the presence of β52Cys in Hb (Fig. 1) significantly increases the resistance of RBCs to attacks by peroxyl radicals (Fig. 2). A previous study showed that two bank vole populations successively colonized Britain at the end of the last glaciation, where they partially replaced one another (Searle et al. 2009). Therefore, there is an area in the north of Britain representing genotypes of the first colonists and in the south genotypes of the second colonists. Interestingly, each colonizing population carried one of the two Hb variants: the first colonists β52Ser and the second colonists β52Cys. Therefore, it was suggested that the higher resistance of their RBCs to oxidative stress could have provided an adaptive advantage to the second colonist over the first colonists and promote the population replacement (Kotlík et al. 2014). Here, we mapped the distribution of the two Hb variants throughout the European distribution of the bank vole to determine the possible geographical origin of the evolutionarily derived β52Cys variant.

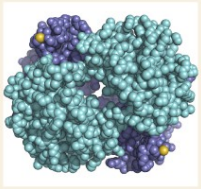


Fig. 1. Model of bank vole haemoglobin showing locations of the β52Cys sulfur atoms on its surface.

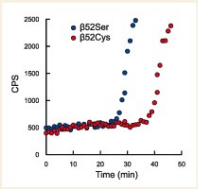


Fig. 2. Bank vole red blood cells (RBCs) containing β52Cys haemoglobin (Hb) significantly delayed the peroxyl-radical induced luminol oxidation (CPS) compared to RBCs containing β52Ser Hb

Pyrosequencing genotyping assays

For genotyping, we used the PyroMark Q24 instrument and software (Qiagen, Valencia, CA, USA), which not only determines the genotype for each individual, but it also enables to quantify the accuracy of the genotype calls. Because the same two amino acids (Ser/Cys) segregate at site β52 in both genes coding for the beta globin chain in the bank vole (Kotlík et al. 2014), we designed two separates assays, one for the major HBB-T1 gene and second for the minor HBB-T2 gene. We used the assays to genotype the site in 518 specimens from 70 regional populations in 23 European countries. Population allele frequencies were calculated with GENEPOP v 4.2 (Rousset 2008). The β52Cys allele frequency data were used as an input for the interpolation of allele frequency surfaces using the Inverse distance weighted (IDW) interpolation method in ArcGIS v 10.2 (ESRI, Redlands, CA, USA).

Geographical pattern and association with Western mtDNA clade

The interpolated allele frequency surfaces reveal a strong geographical pattern of the β52Cys allele in Europe (Fig 3). Only 43 out of the 70 populations carried the β52Cys allele at HBB-T1, of which 12 were fixed for the allele and 31 polymorphic (β52Cys frequency 0.06–0.96). At HBB-T2 the β52Cys shows a more restricted distribution with only six populations fixed for the allele. The distribution of β52Cys at both genes is largely concentrated to the west of Europe, with the majority of the fixed and high-frequency populations in France, Germany and the Netherlands.

Surprisingly, there are two separate increases of the β52Cys frequency at HBB-T1: in the Balkans (Bulgaria) and in southern Scandinavia (Sweden) (Fig. 3).

A comparison of Figs 3 and 4 shows a clear tendency for the β52Cys allele to be associated with the western phylogeographic lineage, as defined by mitochondrial (mt) DNA sequencing (Fig. 4), which most likely originated from woodland refugia in the foothills of the Alps (Kotlík et al., 2006; Magri et al. 2008). Interestingly, β52Cys is absent from the populations of the western mtDNA lineage within the mountain range of the Alps (Fig. 3).

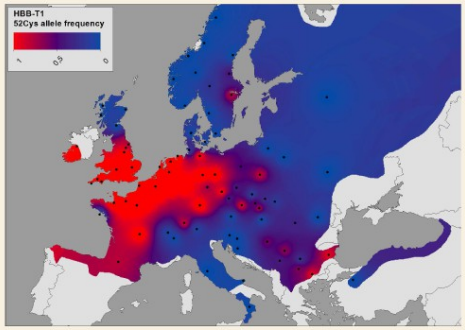


Fig. 3. Geographical distribution of the β52Cys allele at HBB-T1 (top) and HBB-T2 (bottom) shown as interpolated frequency surfaces. Dots represent the centres of the population samples. Data for Britain are from Kotlík et al. (2014).

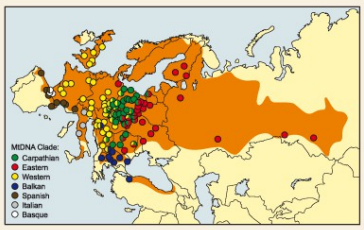


Fig. 4. Distribution of mitochondrial DNA (mtDNA) clades throughout the range of the bank vole. Modified from Wójcik et al. (2010).

Conclusion

The β52Cys Hb likely dispersed post-glacially with the bank voles of the western mtDNA lineage, which were successful in populating large areas in western Europe. The functional advantage of the β52Cys was hypothesised to contribute to the population replacement in Britain and our results suggest it may have been similarly important on the continent. The two separate increases of its frequency challenge a single geographical origin of this functional Hb variant and require further investigation.

References

Kotlík P et al. (2006) Proc Natl Acad Sci USA 103, 14860-14864; Kotlík P et al. (2014) Proc R Soc B 281, 20140021; Magri D et al. (2008) J Biogeogr 35, 450-463; Rossi R et al. (1998) J Biol Chem 273, 19198-19206; Rousset F (2008) Mol Ecol Resources 8, 103-106; Searle JB et al. (2009) Proc R Soc B 276, 4287-4294; Wójcik JM et al. (2010) J Zool 281, 249-262.

Acknowledgements

The work was supported by the Czech Science Foundation (grant number P506-11-1872) and the institutional support (RVO 67985904). Jana Kopecká provided technical assistance.

Annexe 11. Article of P. Kotlík about bank vole haemoglobin polymorphism in Czech popular scientific magazine Živa for which I am part of the co-author team.

Petr Kotlík a kolektiv autorů

Adaptivní fylogeografie: od molekulárních markerů k funkčním genům

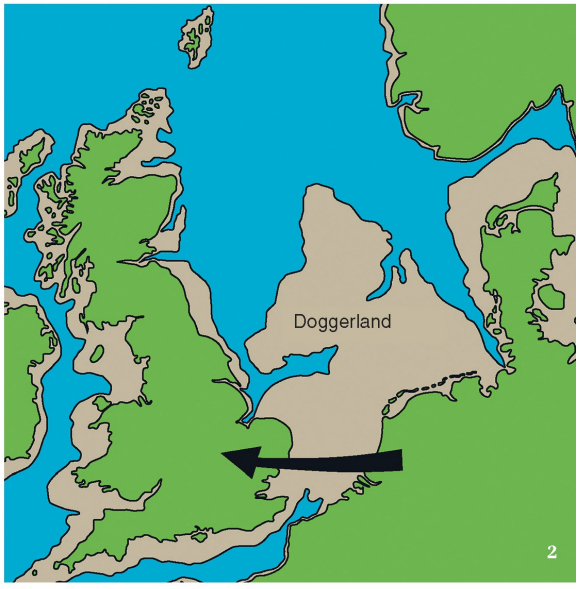
Současné rozšíření rostlin a živočichů mírného pásu je výsledkem kolonizace během oteplení na konci poslední doby ledové přibližně před 10 tisíci lety, kdy druhy postupovaly z glaciálních refugií do dříve neobyvatelných oblastí, především ve vyšších zeměpisných šířkách. Paleontologie přinesla v tomto směru řadu důležitých informací, ale naše současná představa, které populace konkrétních druhů sloužily jako zdroj kolonizace Evropy, vychází zejména z poznatků získaných využitím genetických metod. Hlavní zásluhu má vědní disciplína fylogeografie, studující historické vztahy mezi populacemi na základě porovnání DNA jejich příslušníků. V průběhu posledních 25 let vycházely tyto studie z předpokladu, že genetické vlastnosti dnešních populací odrážejí vlastnosti té populace, která dané území kolonizovala jako první. Nové výsledky však ukazují, že v některých případech byla jedna populace po kolonizaci určité oblasti částečně nebo úplně nahrazena jinou, přicházející z odlišného refugia. Faktorem určujícím, která populace se šířila na úkor jiné, by pak byla schopnost zástupců jednotlivých populací obstát v konkurenci. Populace pocházející z konkrétního refugia tak mohou obývat určitá území ne z čistě geografických důvodů (jako je např. blízkost refugia nebo nepřítomnost migrační bariéry), ale proto, že podmínky a přírodní výběr v daném refugiu upřednostnily vlastnosti přinášející výhodu během kolonizace (např. vyšší reprodukční potenciál, efektivnější termoregulaci nebo odolnost vůči parazitům). Přístup zohledňující takové adaptivní rozdíly mezi populacemi označujeme jako adaptivní fylogeografie.

Vnitrodruhová fylogeografie

Principy fylogeografie poprvé formuloval americký přírodovědec John Avise se svými kolegy. V průkopnické práci zveřejněné v r. 1987 v časopise Annual Reviews in Ecology and Systematics charakterizují fylogeografii jako studium vztahů mezi rodokmenem zástupců stejného druhu a jejich geografickým původem. Vymezují tak nový vědní obor vůči tradiční fylogenetice, jejímž cílem je rekonstrukce vztahů mezi různými druhy, často reprezentovanými pouze jediným zástupcem, jehož geografický původ navíc není rozhodující. Pro fylogeografické studie je naopak podstatná snaha o důkladné pokrytí areálu studovaného druhu, protože pouze zhodnocením příbuznosti jedinců a potažmo populací z různých oblastí – včetně glaciálních refugií a kolonizovaných území – lze získat ucelený obrázek o historii a šíření druhu.

Obvyklý postup zahrnuje určení nukleotidové sekvence vybraného úseku DNA – markeru – zvlášť pro každého jedince a na základě podobnosti sekvencí zrekonstruování příbuzenských vztahů jejich nositelů v podobě genealogického stromu (sekvenování předcházelo určování rozdílů v DNA jednoduššími metodami, především štěpením restrikčními enzymy). Genealogie markeru představuje vlastně ekvivalent fylogenetického stromu, kde ale na konci větví jsou sekvence jedinců stejného druhu (alely genu). Území obývaná blízce příbuznými jedinci, tedy se stejnou nebo velmi podobnou sekvencí DNA, potom byla s největší pravděpodobností kolonizována z jedné zdrojové populace – glaciálního refugia. Oblasti, kde se naopak potkávají geneticky odlišné populace (se značně odlišnými sekvencemi), zpravidla tvoří geografické bariéry nebo zóny, kde se setkaly kolonizační cesty z různých glaciálních refugií. Naprostá většina fylogeografických studií vychází z předpokladu, že faktory určující, které oblasti byly osídleny z kterého glaciálního refugia, nezáležely

1 Ostrov Raasay při západním pobřeží Skotska drobní savci kolonizovali ještě před jeho oddělením od Velké Británie. V pozadí ostrov Skye



2 Na konci doby ledové byla Velká Británie spojena s Evropou pevninským mostem nazývaným Doggerland, který sloužil jako kolonizační cesta. Tání pevninského ledovce mělo za následek vzestup mořské hladiny a přerušení pevninského mostu přibližně před 8,5 tisíci let. Podle: B. J. Coles (1998), upraveno

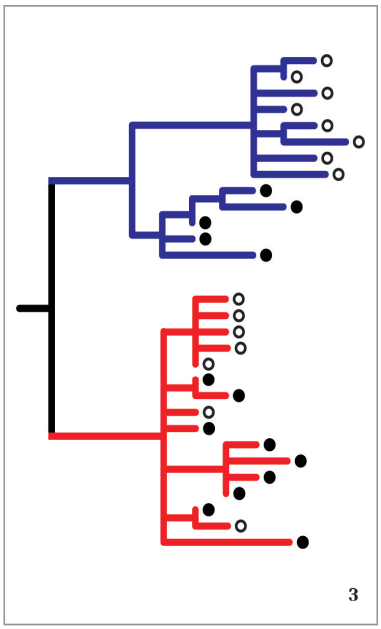
na genetických odlišnostech mezi populacemi. Rozdíly v sekvencích DNA tak závisejí pouze na tom, jestli jejich nositelé pocházejí ze stejné nebo z různých populací, přičemž v druhém případě přibližně odrážejí dobu, po kterou populace oddělovala geografická bariéra. Rozdíly v sekvencích DNA mezi různými zástupci téže populace pak závisejí na velikosti určité populace, jak současné, tak v minulosti. Takové markery, kde nositel jedné sekvence není zvýhodněn přírodním výběrem (selekcí) oproti nositeli jiné sekvence, označujeme jako neutrální. Neutrální markery mohou být úseky DNA neobsahující žádný gen, anebo takové, kde rozdíly mezi sekvencemi nemají vliv na funkci výsledného produktu kódovaného daným genem. Donedávna nebylo příliš důvodů pochybovat, že fylogeografie většiny druhů skutečně neutrální je, a nebyla podstatně ovlivněna přírodním výběrem. Cílem přírodního výběru však může být kterýkoli gen v genomu bez ohledu na to, jaký jsme si zvolili marker. Není proto vyloučené, že mezi populacemi existovaly rozdíly přinášející během kolonizace některým z nich výhodu oproti jiným i v případě, že rozdíly v sekvencích markeru jsou výhradně neutrální.

Keltský lem rejsků a hrabošů ve Velké Británii

Tento pohled zcela změnily výsledky nedávného porovnání fylogeografie drobných savců ve Velké Británii. V práci zveřejněné v r. 2009 v časopise Proceeding of the Royal Society of London jsme s kolegy z Anglie, Skotska a Irska ukázali, že v nejméně pěti případech – u dvou druhů rejsků a tří druhů hrabošů – byla Velká Británie krátce po sobě osídlena dvěma geneticky odlišnými populacemi, z nichž každá pocházela z jiného glaciálního refugia. Velkou Británii během poslední doby ledové pokrýval ledovec a byla součástí evropského kontinentu, s nímž ji spojoval pevninský most v místě dnešního Severního moře – tzv. Doggerland (obr. 2). S postupujícím oteplováním se táním pevninského ledovce začaly zvyšovat hladiny moří, až došlo k zatopení Doggerlandu a přerušení spojení s Británií, která se tak přibližně před 8,5 tisíci let stala ostrovem. Jako marker byla ve čtyřech z těchto pěti případů zvolena mitochondriální DNA (mtDNA), přesněji jeden z genů mtDNA – gen pro protein cytochrom *b*, který je součástí enzymatického komplexu tvořícího článek buněčného dýchacího řetězce (jen u rejska obecného byly porovnány chromozomy). Od samých počátků fylogeografie byla mtDNA jako marker první volbou – v naprosté většině studií v uplynulých 25 letech vědci použili k rekonstrukci vnitrodruhových genealogických vztahů mtDNA.

U všech pěti druhů – rejska obecného a r. malého (*Sorex araneus* a *S. minutus*), hraboše mokřadního (*Microtus agrestis*), hryzce vodního (*Arvicola terrestris*) a norníka rudého (*Clethrionomys glareolus*, obr. 6) – tvoří fylogeografii v Británii dvě hlavní genealogické linie (obr. 3). Jejich rozšíření se navzájem nepřekrývají a populace v okrajových částech Británie – ve Skotsku a jižní Anglii (v případě rejsků i Walesu), patří k jiné linii než populace v centrální části ostrova (obr. 4). Vysvětlením je scénář, kdy v případě každého z druhů byla první příchozí populace v Anglii částečně nahrazena druhou přicházející z jiného refugia a nesoucí odlišnou mtDNA. Blízká podobnost rozšíření okrajových linií rejsků a hrabošů s oblastmi obývanými kulturně a geneticky svébytnými národy hovořícími keltskými jazyky (skotskou gaelštinou, irštinou, velštinou, manštinou a kornštinou) inspirovala označení fylogeografie britských drobných savců termínem Keltský lem.

Fylogeografie mtDNA rejsků a hrabošů ve Velké Británii ukázala, že populace určitého území nemusí nutně být tou populací, která dané území kolonizovala jako první. Vyvolává ale také především otázku, co bylo hnacím motorem, který způsobil, že později příchozí na tak velkém území vytlačili a nahradili ty první. A byla skutečně první populace nahrazena tou druhou, nebo proběhla pouze výměna mtDNA? Ať už šlo o nahrazení populace nebo jen několika genů, jakou úlohu mohl sehrát přírodní výběr upřednostňující vlastnosti zástupců populace pocházející z jednoho glaciálního refugia před zástupci populace z jiného refugia? Najít odpovědi na tyto otázky měl výzkumný projekt, v němž jsme se soustředili na jeden z pěti případů Keltského lemu – na norníka rudého.



3 Genealogické vztahy norníků rudých (*Clethrionomys glareolus*) z Velké Británie (černé kroužky) a z kontinentální Evropy (prázdné kroužky) zrekonstruované porovnáním sekvencí mitochondriální DNA. Norníci v Británii patří ke dvěma genealogickým liniím, které obě přišly z kontinentu. Orig. P. Kotlík

4 až 6 Keltský lem norníka rudého (obr. 6) – jednoho z druhů, které na konci doby ledové kolonizovaly Velkou Británii ve dvou vlnách, přičemž druzí, později příchozí kolonisté nahradili ty první. Blíže v textu. Rozšíření genealogických linií prvních (modře) a druhých (červeně) kolonistů je zobrazeno zvlášť pro mitochondriální DNA (obr. 4) a pro hemoglobin (obr. 5). Přerušované čáry na obr. 5 vyznačují hranice Skotska, Walesu a Cornwallu. Podle: J. B. Searle a kol. (2009), upraveno

7 V kulturní krajině Anglie (zde Cornwall) jsou typickým biotopem norníka zarostlé břehy cest (banks – odtud anglický název norníka bank vole).

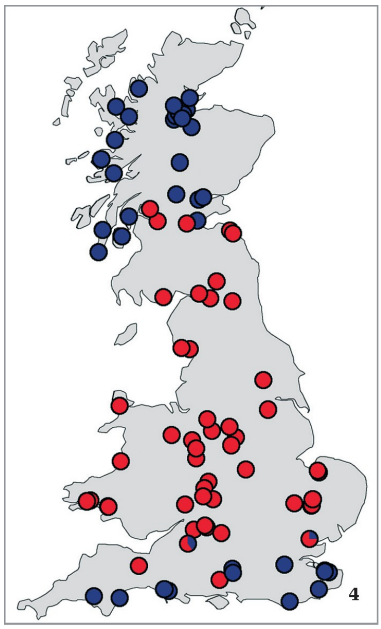
Evoluce mitochondriální DNA norníka rudého

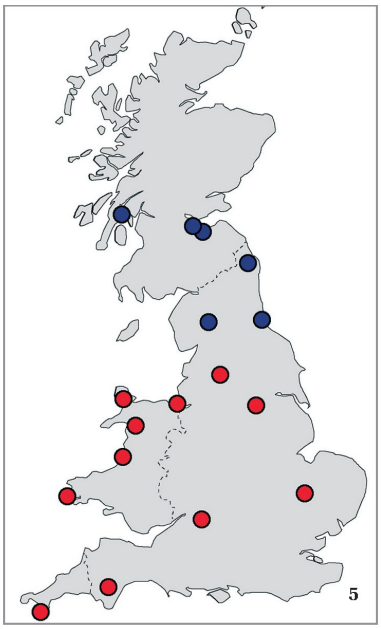
Všech 37 genů obsažených v mtDNA tohoto druhu tvoří jednu vazebnou skupinu, protože mtDNA je předávána mezi generacemi jako celistvá molekula a dědí se jen od matky, aniž by docházelo k mísení s geny od otce (tzv. maternální a klonální dědičnost). To znamená, že pokud by některý z genů byl cílem přírodního výběru, bude jím ovlivněna celá mtDNA. Porovnáním sekvencí kompletního mitochondriálního genomu (více než 16 tisíc nukleotidů) několika zástupců obou britských populací norníka jsme však nezjistili žádné rozdíly, které by poukazovaly na funkční odlišnosti mezi mtDNA prvního a druhého kolonisty. Jejich mtDNA nevykazovala zvýšenou četnost záměn aminokyselin, jež by mohla být důsledkem přírodního výběru. Pouze na jedné pozici jediného genu – pro cytochrom *b* – jsme zjistili významnou změnu fyzikálně-chemických vlastností, konkrétně záměnu aminokyselin alanin za threonin, jejichž postranní řetězce se liší svým objemem (u alaninu je tvořen jedinou metylovou skupinou, zatímco u threoninu jednou hydroxylovou a jednou metylovou skupinou). Přestože tyto rozdíly mohou ovlivňovat funkci enzymatického komplexu, jehož je cytochrom *b* součástí (k záměně došlo v prostoru vazebného místa pro koenzym Q), mutace se vyskytovala v sekvencích zástupců jak prvního, tak druhého kolonisty, a není tak příliš pravděpodobné, že by při kolonizaci přinášela výhodu jedné populaci oproti druhé.

Genomický gradient

Fakt, že všechny geny mtDNA mají stejnou genealogii, také znamená, že pokud chceme zjistit, zda různé části genomu ukazují na stejnou historii nahrazení jedné populace druhou, musíme obrátit pozornost ke genům jaderné DNA. Ta obsahuje mnohonásobně větší počet genů než mtDNA (v genomu laboratorní myši odhadem přibližně 25 tisíc; genom norníka nebyl doposud popsán). Jaderné geny se dědí od matky i od otce, a to nejen nezávisle jeden na druhém, ale také na mtDNA. Bylo tak teoreticky možné, že po příchodu druhých kolonistů do té původní populace pronikla pouze jejich mtDNA a kromě ní žádný, nebo jen několik málo genů jaderné DNA. Nemuselo tedy jít o skutečné nahrazení jedné populace druhou, ale pouze o přenos genů. Vzhledem k vysokému počtu genů v jaderné DNA by nebylo příliš efektivní určit sekvenci jednoho genu po druhém a pro každý rekonstruovat genealogický strom. Na druhou stranu omezit se na jediný, nebo na několik genů by zase neposkytlo reprezentativní obrázek o historii celého genomu.

Zvolili jsme proto trochu jiný přístup, který využívá technologického pokroku


4


5

v sekvenování DNA. Tyto nové technologie, označované jako sekvenování nové generace (Next Generation Sequencing, NGS), umožňují v jednom experimentu získat velké množství sekvencí (obvykle několik milionů) reprezentujících většinu genů z genomu daného druhu, a to hned pro několik různých jedinců najednou (jejich DNA se nejprve individuálně označí). Problém spočívá v tom, že dopředu nevíme, která sekvence pochází z kterého genu. Vstupním materiálem je totiž celková DNA (nebo RNA odpovídající genům aktivním v určitém orgánu, např. slezině, mediátorová RNA, viz dále), navíc z technických důvodů předem fragmentovaná na krátké úseky – kratší než většina genů. Je tedy nejprve nutné vzájemným porovnáním jednotlivých sekvencí určit, které náležejí ke stejnému genu, a z nich potom sekvenci genu sestavit. Tento velmi náročný postup se neobejde bez výkonného počítače. Při použití odpovídající výpočetní kapacity a moderních algoritmů je však docela dobře proveditelný, a tak v dalším kroku mohou být k získaným referenčním sekvencím jednotlivých genů zpětně přiřazeny původní krátké sekvence. Tentokrát jde vždy pouze o ty, které pocházejí z DNA konkrétního jedince, a celý postup se zopakuje pro každého jedince zvlášť. Pečlivým srovnáním sekvencí určitého jedince s referenční sekvencí příslušného genu je pak možné identifikovat rozdíly u daného jedince, i rozdíly mezi jedinci navzájem, a to pro každý z genů. Metody NGS tak umožňují poměrně rychle a s relativně nízkými náklady (v porovnání s tradičním sekvenováním jednoho genu po druhém) porovnat sekvence tisíců genů mezi desítkami jedinců stejného druhu. Takové množství dat posouvá fylogeografii na zcela novou úroveň, protože lze srovnávat populace skutečně v genomickém měřítku.

Vzhledem k množství genů a charakteru dat získaných technikami NGS již není praktické vztahy mezi různými jedinci a populacemi vyjadřovat formou genealogických stromů (pracovali bychom s tisíci různými genealogiemi). Stále častěji se proto uplatňují výpočetní metody, které namísto celých sekvencí genů pracují pouze s polymorfními nukleotidovými pozicemi, tedy těmi, kde se někteří jedinci odlišují od referenční sekvence genu, a tím od ostatních jedinců ze stejné nebo jiné populace. Tyto polymorfní nukleotidové pozice se označují SNP (Single Nucleotide Polymorphism). Pro každého jedince potom můžeme porovnáním jeho genotypů na jednotlivých pozicích SNP s frekvencemi genotypů stejných SNP v různých populacích vypočítat příměs genů z těchto populací v jeho genomu. Tak lze např. v genomu jedinců z různých částí areálu určit podíl genů původem z jednotlivých glaciálních refugií, a tím i hranice mezi oblastmi kolonizovanými z konkrétních refugií, nebo také případy, kdy při kolonizaci došlo k promíchání genů mezi populacemi z více refugií.

V naší genomické studii Keltského lemu norníka rudého jsme tímto postupem zjistili genotypy na více než 10 tisících pozic SNP, z nichž navíc každá leží v jiném genu, a to celkem pro 39 norníků z 6 různých lokalit. Takový počet markerů znamená, že máme možnost udělat si poměrně přesnou představu o podílu genů obou kolonizujících populací v genomu norníků z různých částí Británie. Na obr. 8 je pro každého z 39 norníků modře znázorněn podíl genů v jeho genomu s původem v populaci, která ostrov kolonizovala jako první, a červeně podíl genů z druhé příchozí populace. Na první pohled je zřejmé, že se podíl genů obou populací liší mezi norníky z různých lokalit. Genom jedinců z okolí skotského města Aberdeen, nejseverněji položené lokality v naší studii, neobsahuje žádné zjistitelné stopy po příměsi genů pocházejících z druhé populace. Znamená to, že 100 % jejich genů je původem z populace, která Británii osídlila první. Oproti tomu přes 50 % genomu norníků z okolí Edinburghu na jihu Skotska již tvoří geny druhých kolonistů, a směrem k jihu podíl jejich genomu dále vzrůstá. V genomu norníků z anglických lokalit Doncasteru a Gloucesteru tak již převládají geny v pořadí druhé kolonizující populace, a na nejjižnějších lokalitách v hrabstvích Cornwall a Devon dokonce nenajdeme žádné stopy genomu populace příchozí do Británie jako první. Analýza markerů SNP tak ukazuje zřetelný severojižní gradient v příměsi genů druhé kolonizující populace –


6


7

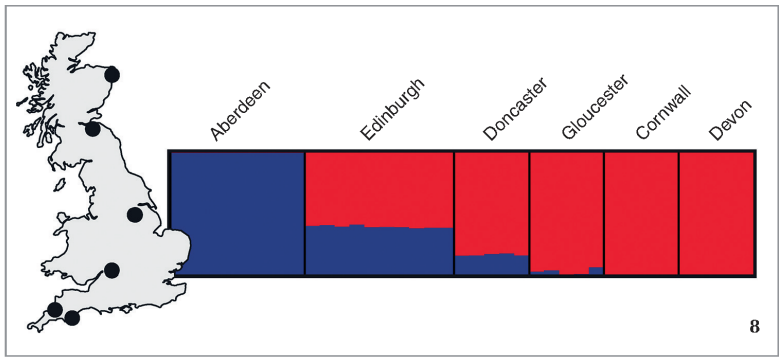
pouze nejsevernější populace si zachovala původní genom, zatímco v obou nejjižněji položených populacích už genom prvního kolonisty zcela nahradily geny druhého příchozího.

Genomická fylogeografie tak potvrzuje scénář, na který poukázala mtDNA, tedy že první příchozí populace norníků byla v Anglii skutečně nahrazena tou druhou, která se sem rozšířila později a přišla z jiného glaciálního refugia. Můžeme vyloučit možnost, že do genomu britských norníků po prvotní kolonizaci pronikla pouze mtDNA nebo maximálně několik málo jaderných genů druhých kolonistů. Fakt, že jde o skutečné nahrazení jedné populace druhou, nezbytně vyvolává otázku, jakou úlohu mohl sehrát přírodní výběr upřednostňující vlastnosti populace pocházející z jednoho glaciálního refugia před populací z jiného refugia. Protože již víme, že samotná mtDNA zdrojem selekční výhody nebyla, je třeba hledat funkční rozdíly zakódované v jaderném genomu. V tomto směru pomohla náhoda, která nás přivedla na stopu fyziologické odlišnosti norníků druhé kolonizující populace, která je za určitých podmínek mohla učinit z hlediska přírodního výběru úspěšnějšími.

Rozdíly v hemoglobinu

V r. 1979 publikoval student univerzity v Cambridgi Stephen Hall vědeckou práci v časopise Journal of Zoology, ve které ukázal, že norníci ze Skotska a severní Anglie mají odlišný hemoglobin než norníci z jihu Británie (obr. 5). S. Hall, dnes emeritní profesor na univerzitě v Lincolnu v Lincolnshiru ve východní Anglii, tehdy porovnal pohyblivost hemoglobinu u zvířat z různých částí země, od Cornwallu po Edinburgh, metodou gelové elektroforézy, která oddělí varianty proteinů lišící se celkovým nábojem (součtem záporných a kladných nábojů jednotlivých aminokyselin, z nichž se protein skládá). Zjistil, že norníci z lokalit jižně od města Doncaster nesli elektronegativnější (více záporně nabitý, a tudíž při elektroforéze rychleji se pohybující) hemoglobin (HbF, z anglického fast – rychlý), než byl hemoglobin norníků ze severu (HbS, slow – pomalý). Pozoruhodné bylo, že se na žádné z lokalit v celé Británii nevyskytovaly oba typy hemoglobinu současně.

Přestože gelová elektroforéza neposkytla žádné informace o funkčních odlišnostech HbS a HbF, vyhraněná geografická rozšíření obou, strukturně zjevně odlišných typů hemoglobinu, vedla S. Halla k vyslovení domněnky, že rozdíly mezi populacemi na severu a na jihu Británie jsou výsledkem působení přírodního výběru, který z nějakého důvodu upřednostňuje HbS na severu a HbF na jihu. Kromě funkčních rozdílů mezi oběma typy hemoglobinu, a tím i možného cíle přírodního výběru, zůstal nevyjasněný původ tohoto polymorfismu – že Velkou Británii postupně kolonizovaly dvě populace norníka, jsme popsali o 30 let později. Stephen Hall ještě pracoval s hypotézou, že jedna z variant vznikla v Británii z té druhé (HbS z HbF, nebo opačně). My dnes víme, že norníci nesoucí HbF jsou potomky druhých kolonistů, kteří si HbF přinesli z kontinentu a na jihu ostrova vytlačili a nahradili ty první, nesoucí HbS. Zajímalo nás proto, jaké rozdíly v molekulární struktuře hemoglobinu jsou zodpovědné za elektroforetické rozdíly mezi HbS a HbF a jaký význam mají z hlediska funkce a potažmo přírodního výběru.



8 Genomický gradient. Graf znázorňuje pro každého z 39 norníků (svislé sloupce) modrou barvou podíl genů v jeho genomu majících původ v populaci, která Velkou Británii kolonizovala jako první; červená barva znamená podíl genů původem z druhé příchozí populace. Pouze nejsevernější populace z okolí skotského města Aberdeen si zachovala genom prvních kolonistů bez příměsi, zatímco v obou nejjižněji položených populacích, v anglických hrabstvích Cornwall a Devon, jsou geny prvního kolonisty již zcela nahrazeny geny druhého kolonisty. Orig. P. Kotlík

9 Model hemoglobinu HbF. Atomy síry (žlutě) cysteinů v podjednotkách beta jsou exponované na povrchu molekuly (obr. a). Hemoglobin HbF od HbS odlišuje záměna jediné aminokyseliny způsobená mutací kodonu pro serin (TCC) na kodon pro cystein (TGC, obr. b). Aminokyseliny serin a cystein se liší pouze v jediném atomu – přítomností síry (S) v cysteinu na místě, kde se v serinu nachází kyslík (O, obr. c). Podle: P. Kotlík a kol. (2014), upraveno

Izolace genů pro hemoglobin

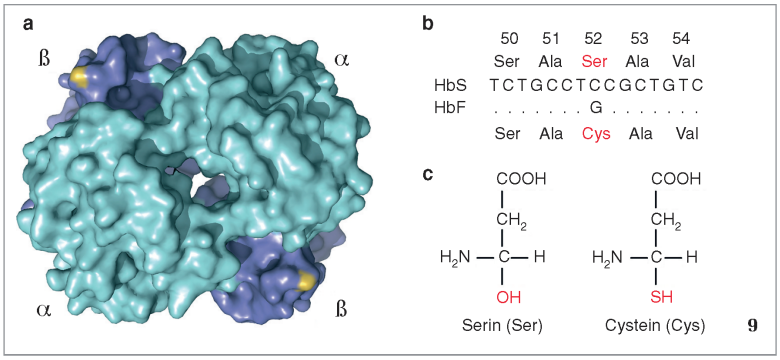
Hemoglobin je protein tvořený čtyřmi polypeptidovými řetězci (globinovými podjednotkami), dva jsou typu alfa a dva typu beta. Jde tedy o tetramerní protein, jehož molekulární strukturu primárně určuje složení jednotlivých podjednotek, přesněji jejich aminokyselinové sekvence. Klíčem k poznání rozdílů struktury HbS a HbF proto bylo určení rozdílů sekvencí genů kódujících jednotlivé globinové podjednotky. Hemoglobin savců je až na výjimky kódován čtyřmi geny – jedním párem genů pro podjednotky alfa a druhým pro podjednotky beta. Protože sekvence žádného z globinových genů norníka nebyla známa, stáli jsme před úkolem tyto geny nejprve izolovat. Předtím jsme však množinu cílových genů zúžili porovnáním elektrostatického náboje podjednotek alfa a beta HbS a HbF. Molekuly obou typů hemoglobinu jsme chemicky rozložili na podjednotky, které jsme od sebe oddělili elektroforézou. Tím se podařilo dokázat, že za rozdíly mezi HbS a HbF zodpovídá vyšší elektronegativita podjednotky beta HbF – podjednotky alfa obou typů hemoglobinů vykazovaly shodnou pohyblivost.

V následujícím kroku jsme proto přistoupili k izolaci a určení nukleotidové sekvence genů kódujících globinové podjednotky beta. Použili jsme k tomu metodu RACE (Rapid Amplification of cDNA Ends), která umožňuje izolovat vybrané geny, aniž známe jejich sekvenci. Stačí znát pouze sekvenci krátké oblasti uprostřed genu, což nepředstavovalo velký problém, protože některé úseky genů pro podjednotky beta klíčové z hlediska funkce jsou shodné mezi různými druhy savců. Výchozím materiálem byla mediátorová RNA (mRNA) získaná ze sleziny norníků, která u hlodavců patří mezi krvetvorné orgány. Při použití mRNA máme jistotu, že jsme získali jen skutečně funkční geny (nefunkční geny netvoří mRNA). Metodou RACE jsme izolovali oba geny kódující podjednotky beta u norníka, označované HBB-T1 a HBB-T2. Nukleotidové sekvence obou genů jsme potom porovnali mezi norníky z různých lokalit v Británii, abychom určili, který z genů a jaké konkrétní mutace zodpovídají za rozdíl mezi HbS a HbF.

Mutace odlišující HbS a HbF

Vzhledem k rozdílnému elektrostatickému náboji HbS a HbF bylo zřejmé, že alespoň jeden ze dvou genů pro podjednotku beta bude kódovat dva různé proteiny odlišující se záměnou nejméně jedné aminokyseliny. K našemu překvapení jsme zjistili, že HbF se od HbS liší právě jedinou nesynonymní mutací (mutací v DNA kódující jinou aminokyselinu) v genu HBB-T1, a co víc, v genu HBB-T2 britští norníci nesou opět právě jen tu stejnou nesynonymní mutaci. Přitom tým vědců z univerzity v Nebrasce vedený Jayem Storzem např. ukázal, že varianty hemoglobinu severoamerického křečka dlouhoocasého (*Peromyscus maniculatus*) přizpůsobené různým nadmořským výškám se liší pěti aminokyselinami v podjednotce alfa a čtyřmi v podjednotce beta. Fakt, že se hemoglobiny HbS a HbF norníka odlišují záměnou jediné aminokyseliny, přináší velkou výhodu při hledání podstaty funkčních rozdílů proti zmíněnému příkladu hemoglobinu křečka, kde bylo k určení funkčních dopadů záměn jednotlivých aminokyselin potřeba použít proteinového inženýrství (k vytvoření všech možných v přírodě se nevyskytujících kombinací jednotlivých aminokyselin).

Mutací, kterou jsme objevili v podjednotce beta hemoglobinu norníka, byla záměna aminokyseliny serin (Ser) na pozici



52 proteinové sekvence za cystein (Cys). Cystein může za určitých podmínek nést záporný náboj (viz dále) a způsobovat tak vyšší elektronegativitu proteinu oproti variantě obsahující serin, který je nenabitý, což by vysvětlovalo rozdíly HbS a HbF při elektroforéze. Kdybychom ovšem nezjistili, že někteří norníci s HbF nesli na pozici 52 v genu HBB-T2 kodon pro serin. Norníci s hemoglobinem HbS přitom opravdu vždy nesli pouze kodon pro serin v obou genech HBB-T1 i HBB-T2, stejně tak u norníků s HbF byl pokaždé kodon pro cystein v HBB-T1. Někteří však měli kodon pro cystein v HBB-T1 a současně pro serin v HBB-T2, přičemž ale elektroforéza jednoznačně určila jejich hemoglobin jako HbF. Jak je možné, že se elektroforézou u těchto norníků nezjistila také přítomnost HbS? Vysvětlení jsme našli v rozdílné míře exprese genů HBB-T1 a HBB-T2 norníka.

Velké množství sekvencí získaných metodou NGS, kde byla výchozím materiálem mRNA jako v naší studii genomického gradientu, lze kromě identifikace SNP použít také k porovnání úrovně exprese různých genů. V případě genů pro podjednotku beta hemoglobinu norníka jsme to provedli tak, že jsme porovnali počty sekvencí odpovídajících genu HBB-T1 s počtem sekvencí původem z genu HBB-T2. Protože jsme mRNA získali ze sleziny, počet sekvencí odpovídajících každému z obou genů odráží množství mRNA od toho genu odvozené a přítomné v krvetvorné tkáni, a může tak sloužit jako měřítko úrovně exprese jednotlivých genů. Tímto způsobem jsme prokázali, že u norníka je gen HBB-T1 více než 20× více exprimovaný než gen HBB-T2. Je to tedy gen HBB-T1, který se rozhodující měrou podílí na produkci mRNA, a tím na syntéze podjednotky beta hemoglobinu u norníka. Ani u jiných savců není neobvyklé, že jeden z dvojice globinových genů je tzv. majoritní gen, zodpovědný za syntézu převážné většiny proteinu, a druhý, minoritní gen se na syntéze podílí relativně velice málo. V případě podjednotky beta u norníka výrazné rozdíly v genové expresi vysvětlují, proč se přítomnost kodonu pro serin v minoritním genu HBB-T2 neprojeví na vlastnostech hemoglobinu norníků s HbF nesoucích kodon pro cystein v majoritním genu HBB-T1, ale kodon pro serin v HBB-T2.

Funkční rozdíly

Ukázali jsme, že hemoglobin HbF od HbS odlišuje záměna jediné aminokyseliny. Existuje-li tedy mezi oběma typy hemoglobinu nějaký funkční rozdíl, musí být způsoben právě touto záměnou serinu za cystein na pozici 52 podjednotky beta. Na první pohled jsou serin a cystein velmi podobné aminokyseliny, které odlišuje jediný atom – přítomnost síry v cysteinu na místě, kde se v serinu nachází kyslík (obr. 9). Rozdíl je však malý jen zdánlivě. Přítomnost sulfhydrylové nebo také thiolové skupiny –SH totiž cysteinu propůjčuje dramaticky odlišné fyzikálně-chemické vlastnosti oproti hydroxylové skupině –OH serinu. Thiol –SH cysteinu totiž může poměrně snadno ionizovat (získat negativní náboj ztrátou vodíkového protonu) a změnit se na thiolátový anion $–S^-$, který za vhodných podmínek představuje vysoce reaktivní funkční skupinu. Takový thiolátový anion potom snadno reaguje s jinou thiolovou skupinou za vzniku disulfidové vazby (kovalentní vazby mezi dvěma atomy síry).

Fyziologicky významné jsou především mezimolekulové disulfidy, které se tvoří mezi thioly na povrchu proteinů a thioly, jež jsou součástí jiných molekul než proteinů – především glutationu (GSH). Jeho molekula se skládá ze tří aminokyselin, z nichž prostřední je cystein. Glutation představuje nejdůležitější antioxidant chránící buňky před oxidačním stresem a poškozením volnými radikály a jinými reaktivními metabolity kyslíku, případně dusíku. V buňkách se vyskytuje ve dvou základních formách, redukované GSH a oxidované GSSG, která není ničím jiným než disulfidem ze dvou molekul GSH spojených disulfidovou vazbou. Disulfid glutationu vzniká reakcí GSH s reaktivními metabolity, kdy je thiol cysteinu v GSH oxidován a reaktivní metabolit naopak redukován – např. peroxid vodíku (H_2O_2) je reakcí se dvěma molekulami GSH redukován na dvě molekuly vody (H_2O), přičemž vzniká jedna molekula GSSG. Pro zachování oxidačně-redukční rovnováhy v buňce musí tedy následovat regenerace GSSG zpět na GSH, které se standardně dosáhne enzymaticky, redukcí glutation reduktázou.

Existují však důkazy, že reaktivní thiol cysteinu přítomný v některých typech hemoglobinu myši se může regenerace GSH účastnit. Děje se tak pravděpodobně prostřednictvím reakce hemoglobinu (HbSH) s GSSG, kdy z molekuly GSSG vzniká molekula glutationylovaného hemoglobinu HbSSG a molekula GSH. Vzhledem k vysoké koncentraci hemoglobinu v červených krvinkách, která je několikanásobně vyšší než koncentrace GSH, tedy může hemoglobin kromě své úlohy přenašeče kyslíku hrát významnou roli v detoxikaci reaktivních metabolitů. Ovšem pouze za předpokladu, že obsahuje cystein s reaktivní thiolovou skupinou. Hemoglobin většiny savců však žádný reaktivní cystein neobsahuje, což platí také pro HbS norníka. Jestli je cystein přítomný navíc v HbF takovým reaktivním cysteinem, jsme zjišťovali sestrojením trojrozměrného molekulárního modelu.

Reaktivita cysteinu, resp. jeho thiolu, závisí jednak na jeho poloze v molekule – jen thiol cysteinu ležícího na povrchu proteinu se dostane do natolik těsného kontaktu s jinými molekulami, aby se mezi nimi mohla uskutečnit chemická reakce. Ne každý takto exponovaný cystein však lze automaticky považovat za reaktivní – jako určující faktor vystupuje rovněž ionizační stav thiolu. Pouze cystein přítomný v buňce ve formě thiolátového aniontu $–S^-$ je reaktivní a jeho reaktivita záleží na poměrném zastoupení aniontové formy. Tu určuje disociační konstanta pK_a thiolu, vyjadřující pH, při němž jsou koncentrace nenabitého thiolu –SH a koncentrace thiolátového aniontu $–S^-$ shodné. Čím nižší je potom pK_a konkrétního thiolu, tím vyšší zastoupení má reaktivní aniontová forma při konkrétním pH. Např. při fyziologickém pH okolo 7,3 bude cystein s pK_a výrazně vyšší než 7,3 v buňce přítomen především jako nenabitý thiol. Naopak cystein s pK_a výrazně nižší než fyziologické pH bude v buňce hlavně ve své aniontové formě.

Náš trojrozměrný model hemoglobinu HbF norníka jednoznačně ukázal, že cystein na pozici 52 podjednotky beta leží na povrchu molekuly HbF (obr. 9a). Na základě stejného modelu jsme potom vypočítali, že elektrostatické interakce atomu síry cysteinu s okolními atomy, jako např. vodíková vazba na kyslík serinu na pozici 50, pomáhají cystein stabilizovat v jeho aniontové formě, a významně tak snižují jeho disociační konstantu pK_a. Ta nezávisle na použité metodě výpočtu leží hluboko pod hodnotou fyziologického pH 7,3 (v blízkosti hodnot 5–6). Takové hodnoty disociační konstanty cysteinu znamenají, že převážná většina jeho buněčné populace (podle našich výpočtů až 94 %) bude tvořena aniontovou formou $–S^-$, což z cysteinu na pozici 52 podjednotky beta HbF činí vysoce reaktivní funkční skupinu.

Reaktivitu cysteinu (schopnost tvořit mezimolekulové disulfidy) jsme následně otestovali. V laboratorních podmínkách lze vzájemnou reakci thiolů katalyzovat oxidačním činidlem. V případě hemoglobinu může oxidace thiolů mít za následek tvorbu polymerů hemoglobinu, a to v případě, že se cystein nachází na povrchu molekuly a jeho thiol je reaktivní. Nechali jsme proto na vzorky krve (hemolyzáty) norníků nesoucích HbF a norníků s HbS působit oxidační činidlo (diamid) a detekovali tvorbu polymerů pomocí elektroforézy. Výsledek jsme porovnali se vzorky, ke kterým bylo přidáno navíc ještě redukční činidlo (merkaptoetanol). Na rozdíl od HbS vykazoval HbF po oxidaci jasné známky tvorby polymerů. Po redukci se však vzorky hemoglobinu HbS i HbF nelišily od příslušných kontrolních vzorků, což potvrzuje, že polymeraci HbF způsobily disulfidové vazby – obdobně se chovají

např. polymerující hemoglobiny myši obsahující reaktivní cystein.

Detoxikace volných radikálů

Pokud hemoglobin sehrál nějakou úlohu při nahrazení jedné populace britských norníků druhou, měly by funkční vlastnosti HbF svým nositelům přinášet výhodu. Zajímalo nás proto, jestli můžeme zjistit rozdíly ve schopnosti červených krvinek obsahujících HbS a HbF odolávat oxidačnímu stresu. Předpokládali jsme, že pokud HbF díky přítomnosti reaktivního cysteinu hraje roli v detoxikaci volných radikálů, budou červené krvinky norníků nesoucích HbF odolávat oxidačnímu poškození lépe než krvinky norníků s HbS.

V pokusu provedeném ve spolupráci s Pavlem Hyršlem a Liborem Vojtkem z Masarykovy univerzity v Brně, jehož výsledky jsme v r. 2014 zveřejnili v časopise Proceeding of the Royal Society of London, jsme k tomu použili luminometrickou metodu pro měření celkové antioxidační kapacity označovanou zkratkou TRAP (Total Radical-trapping Antioxidant Potential). Tato metoda využívá luminiscence (světélkování) oxidačního produktu činidla luminolu k porovnání doby, po kterou jsou různé vzorky schopny vychytávat volné radikály. Signál, který luminol po své oxidaci volnými radikály vydává, lze měřit luminometrem, přičemž náhlý prudký nárůst luminiscence znamená, že právě došlo k vyčerpání kapacity antioxidantů v daném vzorku. Proto čím delší doba uplyne mezi vystavením vzorku působení volných radikálů a nástupem chemiluminiscence, tím větší je antioxidační kapacita vzorku. Červené krvinky norníků nesoucích HbF v našem experimentu vykazovaly trojnásobnou průměrnou hodnotu antioxidační kapacity (vztaženou ke standardu) než u norníků s HbS. Rozdíl byl nejen statisticky průkazný, ale všech 10 norníků s HbF zahrnutých v experimentu mělo antioxidační kapacitu vyšší než kterýkoli z 10 norníků s HbS. Vystavením červených krvinek oxidačnímu stresu se nám tak podařilo ukázat, že přítomnost reaktivního thiolu v hemoglobinu norníků průkazně zvyšuje schopnost jejich červených krvinek odolávat oxidačnímu poškození účinkem volných radikálů kyslíku.

10 Pobřeží Skotska je protkáno řadou zálivů (ve skotské gaelštině označovaných jako lochs) zařezávajících se hluboko do pevniny, které tvoří přirozené migrační překážky. Loch Leven v oblasti Skotské vysočiny. Snímky P. Kotlíka

Adaptivní fylogeografie

Jaký vliv ale mohl takový rozdíl ve fyziologii červených krvinek mít na relativní úspěšnost populací kolonizujících Británii v různou dobu a z různých glaciálních refugií? Fakt, že HbF zvyšuje odolnost červených krvinek vůči oxidačnímu stresu, znamená, že nositelé HbF budou ve výhodě za situace zvýšené potřeby antioxidantů. Produkce volných radikálů a jiných reaktivních metabolitů kyslíku a dusíku významně vzrůstá během energeticky náročných fyziologických stavů, jako je zvýšená fyzická aktivita a investice do reprodukce, nebo při teplotním stresu či parazitární infekci. Není těžké si např. představit, že pokud jsou samice vychovávající větší počet potomků vystaveny vyššímu oxidačnímu stresu, bude mít populace s geneticky podmíněnou lepší schopností detoxikovat volné radikály vyšší reprodukční potenciál než jiné populace stejného druhu.

Fylogeografické studie dlouho předpokládaly, že populace obývající konkrétní území jsou populacemi, které ta území kolonizovaly jako první. Naše studie norníka rudého však ukazuje, že v některých případech (a není důvod, abychom se domnívali, že jde jen o výjimky) byla jedna populace po kolonizaci určité oblasti částečně nebo úplně nahrazena jinou, přicházející z odlišného refugia. Faktorem určujícím, která populace se rozšířila na úkor jiné a která naopak byla nahrazena, by potom byla schopnost zástupců jednotlivých populací obstát v konkurenci. Populace pocházející z určitého refugia tak mohou obývat určitá území ne z čistě geografických důvodů (jako blízkost refugia nebo nepřítomnost migrační bariéry), ale proto, že podmínky a přírodní výběr působící v daném refugiu upřednostnily vlastnosti přinášející výhodu během kolonizace (vyšší reprodukční potenciál, účinnější termoregulaci nebo odolnost vůči parazitům).

Adaptivní fylogeografie tedy znamená snahu porozumět úloze přírodního výběru. Naše studie hemoglobinu norníka se na rozdíl od jiných fylogeografických prací neomezila na analýzu neutrálních molekulárních markerů, ale použili jsme integrovaný mezioborový přístup zahrnující evoluční molekulární genetiku, transkriptomiku, molekulové modelování, biochemii proteinů a buněčnou fyziologii. Jsme přesvědčeni, že podobný přístup je třeba aplikovat i v případě jiných druhů. Pokud je hlavním určujícím faktorem postglaciálního rozšíření populací adaptace a selekce, budou mít takové výsledky klíčový význam pro porozumění změnám v rozšíření druhů během současných i budoucích klimatických změn.

Orig. Vladimír Renčín

Vědecko-výzkumnou práci v laboratoři autora, jejíž výsledky článek představuje, podpořila Grantová agentura ČR (projekt P506-11-1872) a Grantová agentura Akademie věd ČR (projekty IAA600450701 a IAA600450901).

Kolektiv spoluautorů: Silvia Marková, Karolína Filipi, Michaela Strážnická a Jeremy B. Searle

Použitou literaturu uvádíme na webové stránce Živy.