

Univerzita Karlova
3. lékařská fakulta

Dizertační práce

Praha, 2019

Ing. Daniela Šimčíková

Univerzita Karlova
3. lékařská fakulta

Dizertační práce

**The role of glycolytic enzymes in the development
of cancer and metabolic disorders**

**Role glykolytických enzymů v rozvoji nádorových
a metabolických onemocnění**

Školitel: RNDr. Petr Heneberg, Ph.D.

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem řádně uvedla a citovala všechny použité prameny a literaturu. Současně prohlašuji, že práce nebyla využita k získání jiného nebo stejného titulu.

Nesouhlasím s trvalým uložením elektronické verze mé práce v databázi systému meziuniverzitního projektu Theses.cz za účelem soustavné kontroly podobnosti kvalifikačních prací.

V Praze, 14. 11. 2019

Daniela Šimčíková

Identifikační záznam:

Šimčíková, Daniela. *Role glykolytických enzymů v rozvoji nádorových a metabolických onemocnění. [The role of glycolytic enzymes in the development of cancer and metabolic disorders]*. Praha, 2019. Počet stran 157, počet příloh 0. Dizertační práce. Univerzita Karlova, 3. lékařská fakulta, II. interní klinika FNKV a 3. LF UK. Školitel: RNDr. Petr Heneberg, Ph.D.

Keywords: cancer metabolism, CRISPR/Cas9, hexokinase, prediction algorithm

Klíčová slova: nádorový metabolismus, CRISPR/Cas9, hexokinasa, predikční algoritmus

ACKNOWLEDGMENTS

First, I thank to my supervisor, Dr. Petr Heneberg, since he brought me back to science, which I gave up under incredible circumstances. I always appreciated that Petr was always willing to share my doubts and did not let me give it up as I wanted many times. In his lab, I became more independent and responsible scientist in many aspects and I am very grateful that he enabled me to improve my skills with huge amount of freedom. Then, I thank to my great medical students, Dominik Gardáš and Tomáš Pelikán, because they motivated me for making our lab and science better and more ambitious. I also thank to current and former members of our lab for friendly and helpful atmosphere, very helpful colleagues from Dr. Lenka Rossmeislová group at our faculty and flow cytometry core facility at the Institute of Molecular Genetics, AS CR.

I also thank to prof. Vladimír Křen and my former colleagues, particularly my enthusiastic friend Mgr. Bára Fliedrová, from the Institute of Microbiology, AS CR, where I spent my previous studies. Prof. Křen gave me some advice, which I appreciate until now. In that time, I met Dr. Pavel Mader who was my great supervisor at the Institute of Organic Chemistry and Biochemistry, AS CR. I appreciated that Pavel advised me and encouraged me to work abroad.

Therefore, I spent wonderful six months at EMBL, Heidelberg, at the group of Dr. Orsolya Barabas under supervision of Dr. Cecilia Zuliani. At EMBL, I realized how fantastic as well as demanding science may be. I thank to Ceci that she taught me how to improve design of experiments and gave me freedom in order to move my project forward. I also thank to other members of Barabas group for friendly and supportive atmosphere as well as members of EMBL core facilities who were very encouraging and helpful.

Most importantly, I thank to my beloved husband Petr and our wonderful kids Esterka and Pét'a who had huge amount of patience with me and supported me in every situation.

In the end, I have to quote Stephen Hawking's advice, which I am trying to follow: 'One, remember to look up at the stars and not down at your feet. Two, never give up work. Work gives you meaning and purpose and life is empty without it. Three, if you are lucky enough to find love, remember it is there and don't throw it away.'

TABLE OF CONTENT

ABSTRACT	1
ABSTRAKT	2
INTRODUCTION	3
Human hexokinases	3
Hexokinases in metabolic diseases	4
Hexokinases in cancer metabolism	6
CRISPR/Cas systems	11
RNA-guided CRISPR/Cas9	12
CRISPR/Cas9 and other genome editing technologies	14
DNA repair induced by CRISPR/Cas9	15
CRISPR/Cas9 in genetic engineering	16
Other CRISPR/Cas systems in genetic engineering	17
CRISPR/Cas systems in therapeutic genome editing	19
AIMS	21
EXPERIMENTAL PROCEDURES AND METHODS	23
Preparation of the recombinant GCK and its variants	23
Preparation of the recombinant HK2	24
Protein concentration assay	24
GCK kinetic measurements	24
pH optimum of GCK and influence of ATP on buffering capacity	25
Prediction methods used for GCK variations	26
Databases of missense variations in genes encoding proteins associated with Mendelian diseases	31
Selection of genes to validate the model	36

Prediction methods – extended analysis	37
GV approach – extended analysis	37
Prediction method REVEL	39
Statistical methods – GCK and prediction methods	40
Statistical methods – GCK	41
Statistical analyses – extended prediction analysis	41
Preparation of hexokinase knockout cell lines using CRISPR/Cas9	43
Restriction analysis of CRISPR/Cas9 clones	44
Western blotting and immunodetection	45
RESULTS	48
Evidence-based tailoring of bioinformatics approaches to optimize methods that predict the effects of nonsynonymous amino acid substitutions in glucokinase	48
Identification of alkaline pH optimum of human glucokinase because of ATP-mediated bias correction in outcomes of enzyme assays	60
First evidence of changes in enzyme kinetics and stability of glucokinase affected by somatic cancer-associated variations	67
Refinement of evolutionary medicine predictions based on clinical evidence for the manifestations of Mendelian diseases	75
Isoform-specific roles of HK1 in ovarian cancer cells	118
DISCUSSION	124
CONCLUSION	134
LIST OF ABBREVIATIONS	135
REFERENCES	139

ABSTRACT

In this Ph.D. thesis, we aimed to focus on molecular mechanisms that underlie the roles of hexokinases in health and disease. First, we focused on the molecular basis of *GCK-MODY* and possibilities how to predict effects of variations in genes causing Mendelian disorders in general. We performed *in vitro* experiments on GCK and its variants carrying activating, neutral or inactivating variations. Subsequently, we compared these experimental results with outcomes from the state-of-the-art prediction algorithms with distinct backgrounds. As a result of analyses, we realized that the prediction algorithms commonly suffered from low specificity. Therefore, we suggested a method how to tailor numerical outcomes of these prediction algorithms in order to increase specificity. Furthermore, we determined pH optimum of human GCK and HK2 and investigated the influence of ATP concentrations on buffering capacity of commonly used buffers in hexokinase assays.

In the part concerning the role of HKs in tumorigenesis, we studied *in vitro* somatic cancer-associated variations in GCK, which did not give meaningful evidence for a role of GCK in tumorigenesis, although a subset of somatic cancer-associated variations were activating, thus potentially advantageous for tumors. Therefore, we rather moved to the study of HK1 and HK2, which have been reported as important isoenzymes for cancer cells, on the model of ovarian cancer cell line. We have prepared HK1 and HK2 knockout cell lines using CRISPR/Cas9 system. Afterwards, we studied changes of expression levels of proteins involved in metabolic and signaling pathways. We have observed changes indicating that the HK1 KO cells trigger cell survival and proliferation. Nevertheless, HK2 KO cells remain to be studied in a similar manner and further supportive experiments are about to be conducted in a near future.

ABSTRAKT

V předkládané doktorské práci bylo našim cílem objasnit molekulární mechanismy role hexokinas ve zdraví a nemoci. Nejprve jsme se zabývali molekulární podstatou *GCK-MODY* a možnostmi, jak obecně predikovat efekty mutací v genech kódujících Mendelistická onemocnění. Provedli jsme *in vitro* experimenty s *GCK* a jejími variantami nesoucími aktivační, neutrální a inaktivační mutace. Následně jsme porovnali výsledky experimentů s výstupy z nejmodernějších predikčních algoritmů, které mají rozdílný základ. Díky analýzám jsme zjistili, že predikční algoritmy obecně trpí nízkou specificitou. Proto jsme navrhli metodu, jak upravit numerické výstupy predikčních algoritmů, aby se zvýšila specificita. Navíc jsme určili pH optimum lidské *GCK* a *HK2* a zkoumali jsme vliv koncentrací ATP na pufrovací kapacitu pufrů běžně používaných v hexokinasových stanoveních.

V části týkající se role hexokinas ve vzniku a rozvoji nádorů jsme studovali *in vitro* somatické mutace *GCK* nalezené v nádorech. Ačkoliv část těchto mutací byla aktivačních, a tedy potenciálně výhodných pro nádory, studie nepřinesla významnou evidenci role *GCK* pro vznik nádorů. Raději jsme se tedy posunuli ke studii *HK1* a *HK2* na modelu ovariální nádorové linie. U *HK1* a *HK2* bylo již uvedeno, že jsou důležitými isoenzymy pro nádorové buňky. Připravili jsme buněčné linie neexprimující *HK1* a *HK2* metodou CRISPR/Cas9. Poté jsme zkoumali změny úrovně exprese proteinů z metabolických i signálních drah. V nádorových buňkách neexprimujících *HK1* jsme pozorovali změny napomáhající zvýšenému přežívání a proliferaci buněk. Nicméně stále zbývají k podobnému prostudování nádorové buňky neexprimující *HK2* a naše současné výsledky musíme podpořit dalšími experimenty.

INTRODUCTION

Human hexokinases

Hexokinases, enzymes catalyzing the first irreversible step of glycolysis, are present across species. They phosphorylate glucose to glucose 6-phosphate where ATP or ADP are donors of phosphate group. The fate of glucose 6-phosphate is tissue-specific and depends on metabolic demands of the cell, so it can serve primarily for ATP/energy production through glycolysis, biosynthesis through pentose-phosphate pathway or energy storage in glycogen.

Four isoenzymes of ATP-dependent hexokinases, HK1-4, are expressed by mammalian cells, although their presence and expression levels differ among tissues (*Wilson, 2003*). HK1-3 are structurally similar hexokinases with much higher affinity to glucose (K_M about 0.02 mM) compared to HK4, primarily known as glucokinase (GCK), which serves for maintenance of the physiological blood level of glucose (K_M about 5 mM). According to the sequence analysis, GCK (50 kDa) is supposed to be the ancestral hexokinase and other hexokinases (100 kDa) arisen from duplication of its gene (*Tsai & Wilson, 1997; Aderhali et al., 1999*). HK1-3 are assembled from one polypeptide chain into two sequentially homologous domains connected by an α -helix. Unlike HK1 and HK3 with the catalytically active C-terminal domain, HK2 has both domains catalytically active (*Tsai & Wilson, 1997*).

HK1-3 and GCK follow distinct enzyme kinetics in the course of glucose phosphorylation. HK1-3 proceed their reactions according to the Michaelis-Menten kinetics, whereas GCK follows the Hill cooperativity kinetics (*Tsai & Wilson, 1997; Davis et al., 1999*). Unlike GCK, HK1-3 are inhibited by the product, glucose 6-phosphate. Furthermore, HK2 and HK3 are also inhibited by inorganic phosphate, whereas inhibition of HK1 is antagonized by inorganic phosphate. The inhibition by inorganic phosphate manifests independently on effects of glucose 6-phosphate (*Tsai & Wilson, 1997; Aleshin et al., 1998; Aleshin et al., 2000*).

HK1 and HK2 are mostly localized on the outer mitochondrial membrane, HK3 is in a perinuclear compartment (*Wilson, 2003*) and GCK is in the cytosol. In the liver, GCK is regulated by the interaction with the glucokinase regulatory protein (GKRP), which acts as a competitive inhibitor of glucose binding to GCK. This interaction is supported by fructose 6-phosphate and suppressed by fructose 1-phosphate. The complex GCK-GKRP is recruited into the nucleus, until the glucose concentration is elevated; then the complex dissociates and GCK returns to the cytosol (*Beck & Miller, 2013*). HK2 can translocate between mitochondria and the cytosol depending on glucose, glucose 6-phosphate and PKB/Akt, regardless ATP, whereas HK1 remains bound to mitochondria. Consistent with the above-described differences in their localization, HK1 mainly promotes glycolysis, whereas HK2 is involved in both glycolysis and glycogen synthesis (*John et al., 2011*).

Hexokinases in metabolic diseases

Variations in human hexokinase genes can cause disorders of various severity, depending on heterozygous or homozygous manifestation of the variations as well as on the functionality of the transcribed protein. Besides the effects of germline variations in HK1 and GCK (see below), the aberrant expression and activity of HK2 has a prominent role in cancer metabolism (see the chapter Hexokinases in cancer metabolism).

HK1 is a key enzyme in red blood cells, since they rely on the glycolytic pathway providing them energy. Disruptions and deleterious variations in *HK1* genes can lead to non-spherocytic hemolytic anemia (NSHA), autosomal recessive Russe type hereditary motor and sensory neuropathy (Charcot-Marie-Tooth disease type 4G), or autosomal dominant retinitis pigmentosa. Non-spherocytic hemolytic anemia is characterized by severe, chronic hemolysis manifesting in the infancy. NSHA is inherited in an autosomal recessive manner (*Paglia et al., 1981; Rijksen et al., 1983; de Vooght et al., 2009*). Charcot-Marie-Tooth disease type 4G

(CMT4G) is characterized by the onset during early childhood. The CMT4G patients suffer from progressive distal muscle weakness and atrophy, delayed motor development, foot and hand deformities. CMT4G affects particularly the members of the Gypsy ethnic (*Jerath & Shy, 2015*).

The third HK1-associated disease, retinitis pigmentosa (RP), is a dystrophic disorder of the retina causing profound loss of vision or blindness. The symptoms of RP are night blindness, progressive loss of peripheral vision, leading to complete blindness. Variations in a number of genes can cause RP; in addition to autosomal dominant variations in *HK1* gene, variations in *RP65* gene can also result in RP (*Sullivan et al., 2014; Wang et al., 2014*). Moreover, the gene therapy for RP caused by *RP65* mutations has been approved in the EU and USA (this gene therapy product is known under the commercial name LUXTURNA, produced by Novartis). This gene therapy is based on the adeno-associated virus delivery system. In the future, this treatment strategy could also be promising for RP caused by variations in *HK1* and other genes.

GCK and its variations are associated with monogenic diabetes, since GCK regulates insulin secretion in the pancreatic β -cells. Variations in *GCK* can cause both hyperglycemia and hypoglycemia. Heterozygous inactivating variations in *GCK* result in maturity-onset diabetes of the young (*GCK-MODY*), manifesting with mild hyperglycemia, which is often detected later during life. Until now, hundreds of inactivating variations in *GCK* have been reported. In contrast to mild effects of heterozygous inactivating variations, homozygous inactivation of both *GCK* alleles manifests already at birth as a more severe disorder, termed permanent neonatal diabetes mellitus (PNDM). *GCK* may also be affected by activating variations. These are mostly located in the heterotropic allosteric activator site of GCK and cause hyperinsulinemic hypoglycemia. *GCK-MODY* and GCK-induced hyperinsulinemic hypoglycemia are inherited in an autosomal dominant manner (*Gloyn, 2003; Osbak et al., 2009*).

Hexokinases in cancer metabolism

Glucose is an essential source of cellular energy and serves as a carbon source for anabolic pathways in mammalian cells. Most differentiated cells convert glucose to pyruvate via glycolysis. Afterwards, pyruvate is metabolized via the tricarboxylic acid (TCA) cycle and electron transport chain in mitochondria. In that process, known as ‘oxidative phosphorylation’, pyruvate is oxidized to CO₂ and H₂O. The proton-motive force, generated by electron transport chain, is exploited for ATP synthesis from ADP and inorganic phosphate in the presence of Mg²⁺.

In contrast, many tumor cells prefer the less efficient conversion of glucose to lactate, regardless the presence of oxygen. This specific phenotype was first described by Otto Warburg in the 1920s and is commonly called the ‘Warburg effect’ or ‘aerobic glycolysis’ (*Warburg & Dickens, 1930; Pedersen, 1978*). Aerobic glycolysis produces only two ATPs per glucose molecule, whereas oxidative phosphorylation produces up to 36 ATPs per completely oxidized glucose molecule. Warburg originally pointed out that cancer cells suffer from a mitochondrial defect that results in impaired aerobic respiration. However, subsequent studies reported normal mitochondrial function in most cancer cells (*Fanti et al., 2006; Moreno-Sanchez et al., 2007*).

Some studies showed that ATP may never be limiting in proliferating cells as long as they can be supplied with nutrients in circulating blood (*DeBerardinis et al., 2008*). Nevertheless, ATP-deficient cells often undergo apoptosis (*Vander Heiden et al., 1999*). Signaling pathways can sense ATP concentration within the cell. For instance, adenylate kinases convert two ADPs to one ATP and one AMP, thus the increase of AMP activates AMP-activated protein kinase (AMPK). The AMPK activation depends on the tumor suppressor protein LKB1 and leads to phosphorylation of several proteins, for instance Raptor in mTORC1 or acetyl-CoA carboxylase 1, in order to improve energy status in the cell (*Hardie, 2007*).

Interestingly, some tumors consist of two metabolically different subpopulations of cancer cells that function in symbiosis. The cells in one subpopulation employ the ‘Warburg effect’ to produce and secrete lactate, whereas the cells in the second subpopulation import and utilize lactate as their main energy source. The first subpopulation is considered to reflect more hypoxic conditions than the second one (*Feron, 2009; Kennedy & Dewhirst, 2010*).

The fact that hexokinase is bound to the outer mitochondrial membrane is advantageous for tumor cells, since hexokinase obtains access to newly generated ATP and can escape inhibition by glucose 6-phosphate (*Bustamante et al., 1977*). Furthermore, HK2 was proved to be overexpressed in malignant tumors and interacting with the voltage-dependent anion channel (VDAC) (*Nakashima et al., 1986*). HK2 overexpression appears to be rational because of its high affinity to glucose, both catalytically active domains and the hydrophobic N-terminal domain allowing the binding to the VDAC protein (*Bustamante & Pedersen, 1980; Arora & Pedersen, 1988*).

Proliferating cells have anabolic demands and need to produce a large amount of nucleotides, amino acids, and lipids. Apart from ATP, these cells require acetyl-CoA, NADPH and equivalents of carbon. Most mammalian cell lines in culture catabolize glucose and glutamine, which provide most of the carbon, nitrogen, free energy and reducing agents, such as NAD(P)H, necessary for cell growth and division. NADPH is produced in the pentose phosphate pathway, via the conversion of malate to pyruvate catalysed by malic enzyme and via the conversion of isocitrate to α -ketoglutarate catalysed by isocitrate dehydrogenase 1 (IDH1) (*Vander Heiden et al., 2009*).

Glucose addiction of proliferating cancer cells and their disability to metabolize non-glycolytic energetic substrates can be mediated by the activation of the phosphoinositide 3-kinase (PI3K)/Akt signaling pathway (*Buzzai et al., 2005*). PI3K/Akt signaling stimulates glucose uptake and metabolism in cancer cells and plays a key role in the regulation of cell

growth. PI3K signaling through PKB/Akt regulates expression of glucose transporters, increases glucose conversion by hexokinase and stimulates phosphofructokinase (PFK) expression (*DeBerardinis et al., 2008*).

Interestingly, metabolic enzymes can also contribute to tumorigenesis, since germline variations in the TCA cycle enzymes succinate dehydrogenase, fumarate hydratase, or cytosolic IDH1 activate glucose utilization in some tumors under hypoxic conditions (*Baysal et al., 2000; King et al., 2006; Parsons et al., 2008*). Proliferating cells preferentially express pyruvate kinase M2 (PKM2), which is regulated by tyrosine-phosphorylated proteins; thus, tyrosine kinases are also involved in regulation of glucose metabolism. PKM2 is an isoform of pyruvate kinase with low activity, thereby directing the carbon utilization for either biosynthesis or complete catabolism (*Christofk et al., 2008a; 2008b*). Concerning tyrosine kinases, the prototypical tyrosine kinase *c*-Src was found that could phosphorylate human HK1 at Tyr732 and HK2 at Tyr686, thereby activating them. The phosphorylated HK1 at Tyr732 corresponds to the incidence of metastasis of various tumors, thus the phosphorylated HK1 could serve as a marker for metastasis risk (*Zhang et al., 2017*).

The transcription factor hypoxia-inducible factor (HIF)-1 orchestrates the cellular response to hypoxic conditions. HIF-1 regulates transcription of multiple genes, including *HK2*, thereby resulting in a hypoxia-tolerant state of the cell (*Majmundar et al., 2010*) and contributing to proliferative metabolism (*DeBerardinis et al., 2008*). Another pro-survival effect of HK2 manifests in neurons, in which HK2 interacts with phosphoprotein enriched in astrocytes (PEA15) to inhibit apoptosis under hypoxia (*Mergenthaler et al., 2012*).

The tumor suppressor p53 controls metabolic genes and influences glucose metabolism. p53 induces expression of TIGAR (Tp53-induced glycolysis and apoptosis regulator), which leads to PFK inhibition and directs glucose into the pentose phosphate pathway, thereby facilitates NADPH production (*Bensaad et al., 2006*). This may be a response defending the

cells against oxidative stress, since NADPH reduces glutathione that defends cells against reactive oxygen species (ROS). Moreover, the p53-inducible protein TIGAR acts as fructose-2,6-bisphosphatase, and, under hypoxia, it re-localizes to mitochondria and complexes with HK2, thereby increasing HK2 activity (*Cheung et al., 2012*).

Pro-inflammatory cytokines promote glycolysis in breast cancer cells by upregulation of specific microRNAs, such as miR-155 that upregulates *HK2* by either activation of signal transducer and activator of transcription 3 (STAT3), or repressing *mir-143* that acts as a negative regulator of *HK2* (*Gregersen et al., 2012; Jiang et al., 2012; Fang et al., 2012*).

In colorectal cancer cell lines, HK2 inactivation increased expression of HK1. Silencing of both HK1 and HK2 led to decreased cell viability (*Kudryavtseva et al., 2016*). Overexpression of multiple glycolytic enzymes, including HK2, was observed in primary pancreatic ductal adenocarcinoma (PDAC) patient tumors. PDAC is a *KRAS*-driven cancer with a poor prognosis and a high incidence of metastasis. HK2 was shown as highly expressed in PDAC metastases. Consistently, HK2 knockdown resulted in the decrease of primary tumor growth in cell line xenografts and the lower incidence of lung metastasis (*Anderson et al., 2017*). Using the model of primary *PTEN/TP53* null mouse prostate cancer, the elevated HK2, resulting from the activated PKB/Akt, was shown in an androgen-deprived environment, thus ensuring survival of cancer cells. Consistently, HK2 inhibition in prostate cancer cells caused decreased cell viability. This finding is conflicting with androgen deprivation therapy as the accepted treatment for progressive prostate cancer (*Martin et al., 2017*).

Based on the results of a pan-cancer copy number alteration profiling, glycolytic enzymes, including HK1-3, have been found amplified in patient tumors as well as experimental systems. *HK2* amplification appears to be related to p53 variations, whereas *HK1* and *HK3* amplifications are related to amplifications of the oncogenes *MYC* and *MDM2*, and deletion of the tumor suppressor *CDKN2A*. These alterations have been revealed in a case of

breast invasive carcinoma, lung squamous cell carcinoma, ovarian serous cystadenocarcinoma, and serous uterine corpus endometrial carcinoma (*Graham et al., 2017*). According to a meta-analysis of 1,932 patients from 15 studies, HK2 overexpression has been indicated as a poor prognostic marker for gastric cancer, hepatocellular carcinoma and colorectal cancer, but not for PDAC. Moreover, HK2 overexpression was remarkably correlated with tumor size, positive lymph node metastasis, advanced clinical stage and high levels of α -fetoprotein (*Wu et al., 2017*). Unlike HK2 overexpression in hepatocellular carcinoma, GSK expression is suppressed. In a mouse model of liver tumorigenesis, HK2 deletion decreased the incidence of tumors. Consistently, HK2 knockdown in human hepatocellular carcinoma cells inhibited tumorigenesis and led to cell death. Furthermore, serine and glycine uptake and oxidative phosphorylation were increased, and served as compensatory mechanisms (*DeWaal et al., 2018*).

In liver cancer cells, glycolytic activity is inversely correlated with autophagy level. In this model, HK2 was found to be ubiquitinated at Lys63 by the E3 ligase TRAF6 and further processed by autophagic degradation, when autophagy mechanisms proceeding properly (*Jiao et al., 2018*). Another role of HK2 in autophagy has been studied for its connection with telomerase, a ribonucleoprotein complex of telomerase reverse transcriptase (TERT) and telomerase RNA component (TERC). HK2 inhibition in HepG2 cells suppressed TERT-induced autophagy, since TERT promotes autophagy through an HK2-mTOR pathway, in which HK2 activation silences mTOR activity. Furthermore, telomerase binds to the HK2 promoter through TERC, thereby promoting HK2 expression (*Roh et al., 2018*).

As mentioned above, HK2 is ubiquitinated at Lys63 by TRAF6, but Lys63-linked ubiquitination is also mediated by the HectH9 E3 ligase. Moreover, HectH9 ubiquitinates the p53 tumor suppressor at Lys48, thereby downregulating p53. HectH9 is upregulated upon hypoxia and promotes tumorigenesis (*Bernassola et al., 2008*). In addition to the above-

mentioned report on autophagy (*Jiao et al., 2018*), Lys63-linked ubiquitination enables HK2 to bind to mitochondria and promotes glycolysis (*Lee et al., 2019*).

CRISPR/Cas systems

Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems serve in bacteria and archaea as RNA-navigated adaptive immune systems which protect these organisms against nucleic acids from invading viruses and plasmids (*Wiedenheft et al., 2012*). In these bacterial adaptive immune systems, RNAs complementary to nucleic acids originated from invaders detect and silence foreign nucleic acids. CRISPR/Cas systems consist of *cas* genes organized in operons and CRISPR arrays which target particular sequences in the genome and are interspersed with identical repetitive DNA (*Wiedenheft et al., 2012*). The common components of all CRISPR/Cas systems are the *cas1* and *cas2* genes (*Amitai & Sorek, 2016*). Furthermore, phylogenetic analyses of Cas1 revealed existence of several versions of CRISPR/Cas systems (*Kunin et al., 2007*).

CRISPR/Cas-mediated immunity proceeds in three steps. In the first (adaptive) phase, a short fragment of foreign DNA (the protospacer) is integrated primarily at the leader end of CRISPR locus. Afterwards, in the expression (second) and interference (third) phases, the repeat-spacer sequences are transcribed into a precursor CRISPR RNA (pre-crRNA), which is cleaved enzymatically. The yielded short crRNA can interact with complementary protospacer sequences of invading viral or plasmid targets. The silencing of foreign nucleic acids is processed by Cas proteins which form complexes with the crRNAs (*Haurwitz et al., 2010*; *Deltcheva et al., 2011*).

According to the latest classification, the CRISPR/Cas systems are divided into two distinct classes, based on the characteristics of the effector module. The Class 1 systems include the most common and heterogeneous type I, type III that is more common in archaea than

bacteria, and the rare type IV. The common features of the type I and III CRISPR systems are specialized Cas endonucleases processing the pre-crRNAs and large multi-Cas protein complexes which recognize and cleave foreign nucleic acids based on the mature crRNAs complementarity (*Jackson et al., 2014; Zhao et al., 2014; Koonin et al., 2017*).

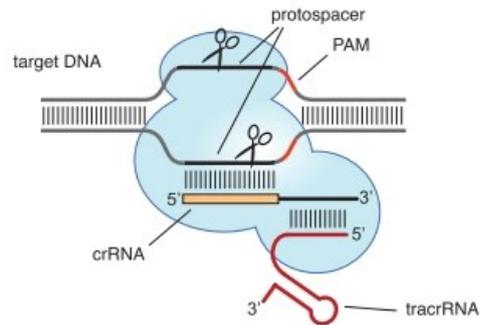
On the other hand, the Class 2 consist of a single, large, multidomain protein and includes three subtypes. The first Class 2 subtype is the well-characterized type II, in which the Cas9 endonuclease is considered the only protein responsible for RNA-guided silencing of foreign DNA (*Sapranauskas et al., 2011*), and the second subtype is represented by the type V with the putative Cpf1 endonuclease. Due to metagenomics analyses, the Class 2 has been completed with the third subtype, type VI, with domains displaying RNase activity, such as Cas13 (*Jinek et al., 2014; Abudayyeh et al., 2017; Koonin et al., 2017*).

RNA-guided CRISPR/Cas9

The first evidence of the CRISPR/Cas employment in RNA-programmable genome editing has been published by *Jinek et al. (2012)*. They proved that the Cas9 protein needs a base-paired structure located between the transactivating crRNA (tracrRNA) and the targeting crRNA in order to cleave the targeted DNA sequence. The cleavage of the Cas9 endonuclease is directed by both complementarity between the crRNA and the target protospacer DNA and a short motif, which is known as the protospacer adjacent motif (PAM) (Fig. 1).

Moreover, the cleavage mechanism of the Cas9 endonuclease has been revealed. To trigger the Cas9-mediated cleavage of plasmid DNA, both the mature crRNA and the *trans*-activating tracrRNA are necessary. The *trans*-activating tracrRNA has two crucial functions – initiating the pre-crRNA processing by the enzyme RNase III (*Deltcheva et al., 2011*) and activating the crRNA-navigated DNA cleavage by Cas9 (*Jinek et al., 2012*).

Cas9 programmed by crRNA:tracrRNA duplex



Cas9 programmed by single chimeric RNA

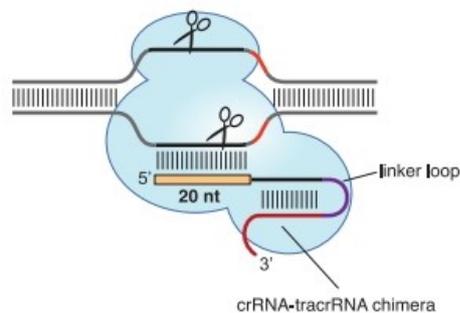


Fig. 1. Scheme of DNA cleavage by the complex assembled from Cas9, tracrRNA and crRNA. In the top, Cas9 is guided by the activating tracrRNA and targeting crRNA. In the bottom, the fused crRNA and tracrRNA, a chimeric RNA is used for the engineered CRISPR/Cas9 (*Jinek et al., 2012*).

The cleavage of plasmid DNA by Cas9 produces blunt ends at a position 3 bp upstream of the PAM sequence. Within short double-stranded DNA duplexes, the DNA strand complementary to the crRNA is cleaved at a site 3 bp upstream of the PAM sequence. In contrast, the non-complementary DNA strand is cleaved at one or more sites within 3 to 8 bp upstream of the PAM. The PAM sequence is recognized specifically by Cas9 as a prerequisite for DNA binding and following strand separation prior to Cas9 cleavage. The PAM sequence and position is varying according to a CRISPR/Cas system type (*Mojica et al., 2009*). The

turnover number of Cas9 is comparable to that of restriction endonucleases and ranges from 0.3 to 1 min⁻¹ (*Jinek et al., 2012*).

Based on the structural study, Cas9 consists of domains homologous to HNH and RuvC endonucleases (*Makarova et al., 2011*). The HNH domain cleaves the complementary DNAs strand, whereas the Cas9 RuvC-like domain cleaves the non-complementary DNA strand (*Jinek et al., 2012*). The endonuclease Cas9 from *Streptococcus pyogenes* (SpCas9) can be navigated by a sgRNA to any genomic locus followed by a 5'-NGG PAM sequence and a 20-nucleotide guide sequence within the sgRNA, responsible for genome targeting, thus SpCas9 can be easily engineered according to a gene of interest (*Jinek et al., 2012*).

CRISPR/Cas9 and other genome editing technologies

Compared to other genome editing technologies, including zinc-finger nuclease (ZFN) (*Miller et al., 2007*) and transcription activator-like effector nucleases (TALENs) (*Hockemeyer et al., 2011; Zhang et al., 2011*), CRISPR/Cas9 represents a system that is significantly easier to design, specific, efficient and suitable for high-throughput gene editing in different cells and organisms. In general, custom ZNFs are difficult to engineer, which severely decreases their potential to become a widespread technology (*Wood et al., 2011*).

In contrast to TALENs, which require the design of two new TALEN genes for every new DNA sequence (*Schmid-Burgk et al., 2013*), Cas9 can be targeted to the desired genome site by simply designed oligonucleotides encoding a 20-nucleotide guide sequence. Given the cleavage pattern, TALENs cleave nonspecifically in the 12-24-bp linker between the pair of TALEN monomer-binding sites (*Miller et al., 2011*), unlike the specific cleavage 3 bp upstream of the PAM sequence by SpCas9 (*Jinek et al., 2012*). Another advantage of SpCas9 stems from the fact that Cas9 can target multiple genome loci simultaneously by delivery a combination of sgRNAs to the cells (*Ran et al., 2013a*).

DNA repair induced by CRISPR/Cas9

The Cas9 endonuclease promotes genome editing by introducing a double-strand break (DSB) into a targeted gene. Then, the cleaved locus undergoes one of two major DNA repair pathways – the error-prone non-homologous end-joining (NHEJ) or the high-fidelity homology-directed repair (HDR).

In mammalian cells, NHEJ is the preferential pathway for DSB repair. In the course of the NHEJ process, DSBs are re-ligated which leads to the formation of insertion/deletion (indel) variations. NHEJ can result in gene knockouts, when indels occur within an exon (*Barnes, 2001*).

An alternative pathway of DNA repair is HDR. HDR require the presence of a repair template, which is provided endogenously during the S and G2 cell cycle phases, since HDR is generally active in dividing cells, or can be introduced exogenously (*van den Bosch et al., 2002*). The exogenous repair template can either be provided in the form of double-stranded DNA with homology arms flanking the insertion sequence, or single-stranded DNA oligonucleotides (*Saleh-Gohari & Helleday, 2004*). In some models, particularly for cancer research, HDR is disadvantaged because of variations occurring in key proteins that are involved in this type of DNA repair. For instance, HDR requires the recruitment of *BRCA* genes that are often mutated in some cancer cell types, to DSB sites.

The choice between NHEJ and HDR is regulated by the 53BP1 protein. This pro-NHEJ factor limits homologous recombination by blocking DNA end resection as well as inhibiting BRCA1 recruitment to DSB sites (*Hustedt & Durocher, 2016*). To increase HDR efficiency in CRISPR/Cas9 editing, *Canny et al. (2017)* developed an inhibitor of 53BP1 based on the ubiquitin structure, since 53BP1 recognizes histone H2 ubiquitylated on Lys15 (*Fradet-Turcotte et al., 2013*). The inhibition of 53BP1 significantly increased efficiency of HDR-based genome editing in human and mouse cells. For the same purpose, the regulation of BRCA1-

PALB2-BRCA2 complex assembly has been suggested promoting HDR during G1 phase of the cell cycle, since the BRCA2 recruitment to DSBs is blocked by the inhibition of BRCA1-PALB2-BRCA2 assembly in G1 cells (*Orthwein et al., 2015*).

CRISPR/Cas9 in genetic engineering

For the practical and widespread use in genome engineering, the human codon-optimized Cas9 from *S. pyogenes* and a chimeric (the fused crRNA and tracrRNA) sgRNA were created (*Jinek et al., 2012; Ran et al., 2013a; Fig. 1*). Furthermore, the engineered Cas9 endonuclease was used for enhancement of genome editing specificity, since specificity of the wild-type Cas9 can be influenced by multiple mismatches between the sgRNA and its complementary target DNA sequence (*Ran et al., 2013b*). This unspecific base pairing can lead to potential off-target DSBs and indel formation (*Fu et al., 2013*). In the improved strategy, the D10A mutant nickase variant of Cas9 (Cas9n) was combined with a pair of sgRNAs complementary to opposite strands of the target site. This applied design relies on the synergistic interaction of two Cas9n, similarly to dimeric ZFN and TALENs, thus minimizing off-target mutagenesis, since individual nicks (single-strand DNA breaks) are predominantly repaired by the high-fidelity base excision repair pathway (*Wood et al., 2011*).

To overcome the limitation of gene targeting in genetic screens, the combinatorial screening using orthogonal Cas9 endonucleases from *Staphylococcus aureus* and *Streptococcus pyogenes* was developed, thereby combining two PAM sequences and increasing the choice of targeted sites (*Najm et al., 2018*).

Based on the CRISPR/Cas9 system, a method called CRISPR interference (CRISPRi) has been developed in order to repress expression of targeted genes. This method employs a catalytically dead Cas9, which lost endonuclease activity due to variations, but is still able to form complexes with sgRNAs. The RNA-guided dCas9 can specifically interfere with

transcription processes and is associated with likely negligible off-target activity (*Qi et al., 2013*). The modified method for silencing was achieved by a fusion of dCas9 with the Krüppel-associated box (KRAB) domain, since KRAB is an effective transcription repressor that recruits a heterochromatin-forming complex that causes histone methylation and deacetylation (*Thakore et al., 2015*).

Other CRISPR/Cas systems in genetic engineering

Zetsche et al. (2015) identified another Cas effector, called Cpf1 (CRISPR from *Prevotella* and *Francisella* 1) that have some distinct features from Cas9 and is useful for genome editing in human cells. This putative Class 2 CRISPR system, a type V CRISPR/Cas system, has been indicated in several bacterial genomes (*Schunder et al., 2013*). Unlike Cas9 systems, Cpf1-associated CRISPR arrays do not require *trans*-activating crRNAs. In contrast to the G-rich PAM sequence for Cas9 systems, Cpf1-crRNA complexes cleave the target DNA based on a T-rich PAM sequence. Cpf1 introduces a staggered DSB with a 4- or 5-nucleotide 5'-overhang. The mechanism and structure of Cpf1 has been revealed a year later (*Fonfara et al., 2016; Yamano et al., 2016*). Unlike Cas9, Cpf1 contains the RuvC domain but lacks a second endonuclease domain, the remaining domains are responsible for the sequential cleavage of the non-target and target strands, thus generating cohesive DSBs.

Two independent studies proved that use of CRISPR/Cpf1 from *Acidoaminococcus* sp. (AsCpf1) and *Lachnospiraceae* bacterium (LbCpf1) results in fewer off-target effects than in the case of Cas9 endonucleases. The first study reported that Cpf1 tolerates single or double mismatches in the 3'PAM-distal region rather than in the 5'PAM-proximal region (*Kim et al., 2016*), concurrently the second study confirmed this finding (*Kleinvister et al., 2016*). Moreover, the use of the preassembled, recombinant AsCfp1 and LbCfp1 ribonucleoproteins has been suggested in order to overcome the off-target activity (*Kim et al., 2016*).

In 2017, *Abudayyeh et al.* reported the Class 2 type VI RNA-guided RNA-targeting CRISPR/Cas effector Cas13a, which can be engineered for RNA knockdown and binding in mammalian cells. Based on the screening of fifteen orthologs, Cas13a from *Leptotrichia wadei* (LwaCas13a) has been determined as the most efficient. LwaCas13a can be heterologously expressed in mammalian and plant cells providing the efficient level of knockdown with significantly higher specificity compared to RNAi. CRISPR/Cas13 has been also proposed as a highly sensitive, rapid and undemanding viral diagnostic platform (*Myhrvold et al., 2018*).

Recently, the third genome-editing platform called CRISPR-CasX has been introduced (*Liu et al., 2019*). CRISPR/CasX was proved to modify genomes of *Escherichia coli* and humans. According to metagenomics analysis of microbial DNA, CasX defended against bacterial transformation by plasmid DNA. Interestingly, CasX shows no sequential similarity to other CRISPR-Cas enzymes, except for the presence of a RuvC nuclease domain. However, the RuvC domain of CasX has less than 16% identity with RuvC domains of either Cas9 or Cas12a. CasX probably evolved from a TnpB-type transposase by an independent insertion into ancestral CRISPR loci. Based on cryo-electron microscopy structures of the CasX, gRNA and double-stranded DNA assembly, an ordered no-target and target-strand cleavage mechanism has been revealed. This mechanism may explain better cleavage mechanisms of CRISPR/Cas enzymes with a single active site, such as Cas12a (*Fonfara et al., 2016; Yamano et al., 2016*). CasX introduces a staggered DSB in DNA at sequences complementary to a 20-nucleotide sequence of its gRNA and adjacent to a TTCN PAM sequence. CasX originated from *Deltaproteobacteria* (DpbCasX) and *Planctomycetes* (PlmCasX) are useful both for genome editing with efficiency comparable with Cas9, and CRISPRi as an engineered deactivated CasX (*Liu et al., 2019*).

CRISPR/Cas systems in therapeutic genome editing

First, *Schwank et al. (2013)* optimized the CRISPR/Cas9 system for the correction of the cystic fibrosis transmembrane conductor receptor (CFTR) locus by HDR in primary adult stem cells derived from cystic fibrosis patients. A year later, the CRISPR/Cas9 editing by HDR was used for the correction a *FAH* (a gene encoding fumarylacetoacetate hydrolase) variation in hepatocytes in a mouse model of the human disease hereditary tyrosinemia. The procedure that led to the expression of the wild-type FAH protein was successful in 1/250 liver cells (*Yin et al., 2014*). In the same year, the exploitation of the CRISPR/Cas9 system for rapid development of mouse liver cancer models was published, when targeting the tumor suppressor *PTEN* and *p53* *in vivo* in wild-type mice (*Xue et al., 2014*). CRISPR-Cas9 genome editing was feasible for the primary open-angle glaucoma treatment, caused by variations in *myocilin* gene in the *in vivo* mouse model (*Jain et al., 2017*).

The breakthrough in genome editing using CRISPR/Cas9 has brought the ability to conduct genome-wide screens in human cells (*Chow et al., 2017; Joung et al., 2017*). Genome engineering of human pluripotent stem cells remained difficult, with lower efficiencies compared to some tumor cell lines or mouse embryonic stem cells. The explanation was given by a study pointing out that DSBs induced by Cas9 trigger a *TP53*-dependent toxic response, thus killing most human pluripotent stem cells. Unfortunately, transient *TP53* inhibition could lead to a higher mutational risk implicating a risk of cancer (*Ihry et al., 2018*).

Regarding expanding possibilities of CRISPR/Cas systems in gene therapy, the stronger attention has been paying on the safety and reliability of potential CRISPR-based therapeutics. Since the on-target and off-target effects vary greatly with individual sgRNAs, many efforts have been made to improve the computational design of sgRNAs with predicted off-target sites in order to minimize the off-target effects and concurrently maximize the on-target activity (*Doench et al., 2016*). Furthermore, the human genome contains many disease-unrelated

genetic variations in every individual patient; these variations might also influence effects caused by Cas endonucleases. By analyses of the Exome Aggregation Consortium and 1000 Genomes Project datasets, the design of patient-specific sgRNAs has been proposed for ensuring the safety of CRISPR-based therapeutics and the pre-therapeutic whole genome sequencing has been suggested (*Scott & Zhang, 2017*).

Deep-sequencing data from 81 genome-editing projects on mouse and rat genomes allowed to predict 1,423 off-target sites and confirm 32 of them, thereby showing that the improved design for CRISPR/Cas9 reduced off-target variation rates (*Anderson et al., 2018*). Furthermore, *Anderson et al. (2018)* studied the impact of genome editing in ten mouse embryos treated with a sgRNA and found 43 off-target effects, 30 of which were predicted. Unfortunately, further explorations the repair of DSBs induced by CRISPR/Cas9 indicated adverse on-target effects, such as large deletions (over many kilobases) and more complex genomic rearrangements (e.g., crossover events), in mouse embryonic stem cells, mouse hematopoietic progenitors and a human differentiated cell line (*Kosicki et al., 2018*).

AIMS

1/ Preparation of the recombinant GCK and its variations, their biochemical characterization and the comparison of experimental results with outcomes from prediction methods (*Šimčíková et al., 2017*).

We aimed to prepare and characterize of GCK variants primarily found in Czech patients suffering from *GCK-MODY*, and measure their kinetic characteristics and stability. We aimed to compare these experimentally obtained results with the outcomes of prediction methods, which are used in personalized medicine. We aimed to test outcomes these prediction methods by the comparison of their outcomes with experimental data on GCK variants that were generated in the present study as well as with those that were previously published.

2/ Determination of pH optimum of human GCK, pH influence on GCK and HK2 and the influence of ATP concentrations on buffering capacity (*Šimčíková & Heneberg, 2019*).

We aimed to determine pH optimum of GCK and investigate pH influence on GCK and HK2 activity. We aimed to investigate the influence of ATP concentrations on buffering capacity of buffers often used in hexokinase activity assays.

3/ Study of GCK activating variations (*Těšínský et al., 2019*).

We aimed to study GCK activating variations, determine their kinetic parameters and temperature stability in the structural context of the GCK molecule.

4/ Tailoring of prediction methods for their use in personalized medicine (*Šimčíková & Heneberg, subm.*).

We aimed to extend the outcomes of Aim #1 to predictions of effects of other genes that are causative for Mendelian diseases. We aimed to build datasets of genes, the variations of which variations cause human inherited diseases and apply the state-of-the-art prediction methods to the datasets analysis. We aimed to improve settings of prediction methods and validate suggested settings.

5/ Characterization of effects of HK1 and HK2 deletions in ovarian cancer cell lines.

We aimed to implement CRISPR/Cas9 technology in order to produce knockout cell lines. As a proof of concept for genome editing experimental design, we aimed to use the HEK293T cell line. To analyse the effects of HK1 and HK2 deletions, we aimed to prepare knock-outs in the ovarian cancer cell line TOV-112D, since ovarian cancer cells, unlike other types of cancers, expresses preferentially the HK1 isoenzyme. We aimed to investigate the adaptation of HK1 knockout cells and changes in expression levels of metabolic enzymes and proteins involved in signaling pathways.

EXPERIMENTAL PROCEDURES AND METHODS

Preparation of the recombinant GCK and its variants (*Šimčíková et al., 2017; Šimčíková & Heneberg, 2019; Těšínský et al., 2019*)

We expressed glucokinase (GCK) from the expression vector, pGEX-5X-2, with an insert that encoded the wild-type GCK isoform 1; the expression vector was provided as a kind gift from Dr. Navas (Universidad Complutense de Madrid) (*Garcia-Herrero et al., 2007*). We introduced variations into the expression construct via site-directed mutagenesis (QuikChange site-directed mutagenesis kit, Agilent Technologies, Santa Clara, CA). We verified all the constructs via bidirectional Sanger sequencing.

We prepared GCK and its mutant forms as fusion proteins with N-terminal glutathione-S-transferase (GST) in the *Escherichia coli* BL21 Gold(DE3) strain (Agilent Technologies, Santa Clara, CA). We grew the cells at 37°C to OD 0.7, and induced the expression by adding IPTG to a final concentration of 0.2 mM. We incubated the culture at 22°C for 16 hours with orbital shaking (240 rpm). Afterwards, we harvested the cells by centrifuging, and resuspended the pellets in a breaking buffer (25-fold smaller volume than the culture volume; PBS, pH 7.4 containing 4 mM MgCl₂, 1 mM PMSF, 5 mM DTT, 0.5% Triton X-100, lysozyme and DNase I) followed by 30 min of incubation at room temperature. We lysed the cells via mild sonication on ice. We centrifuged the lysate (4°C; 20,000×g) and immediately incubated the supernatants with Glutathione Sepharose (GE Healthcare Life Sciences, Chicago, IL). Subsequently, we washed the beads twice and eluted GST-GCK with 50 mM Tris, 200 mM KCl, pH 8.0, containing 5 mM DTT and 10 mM glutathione. We performed the entire purification procedure at 4°C.

Preparation of the recombinant HK2 (*Šimčíková & Heneberg, 2019*)

We introduced the insert encoding HK2 into pET-28a(+) and expressed HK2 in BL21(DE3)pLysS *E. coli*. We induced HK2 expression by the addition of 1 mM IPTG and subsequently cultivated the cells for 16 h at 22 °C. Afterwards, we purified HK2 using HisTrap HP (GE Healthcare, Chicago, IL).

Protein concentration assay (*Šimčíková et al., 2017; Šimčíková & Heneberg, 2019; Těšínský et al., 2019*)

We determined the protein concentration using a Bradford assay (Serva, Heidelberg, Germany) with bovine serum albumin used as a standard.

GCK kinetic measurements (*Šimčíková et al., 2017; Šimčíková & Heneberg, 2019; Těšínský et al., 2019*)

We measured the GCK activity spectrophotometrically using a coupled reaction with glucose-6-phosphate dehydrogenase (Sigma-Aldrich, St. Louis, MO) and determined the increasing concentration of NADPH at 340 nm as described previously (*Liang et al., 1995; Davis et al., 1999*). One unit (U) of GCK was defined as the amount of enzyme that phosphorylated 1 μ mol of glucose per min at 30°C under assay conditions. In the case of glucose as the variable substrate (0–200 mM), we measured these assays using two concentrations of ATP – 0.5 mM and 5 mM; the GCK activity exhibited a sigmoidal dependency, which satisfied the Hill equation. However, the GCK activity with variable ATP concentrations (0–5 mM) followed hyperbolic Michaelis-Menten kinetics. We performed GCK assays with variable ATP concentrations at two glucose concentrations: at the corresponding $S_{0.5}$ and 50 mM. We performed the competitive inhibition with *N*-acetylglucosamine (GlcNAc) at 5 mM glucose and 5 mM ATP under identical assay conditions.

In the study by Šimčíková *et al.* (2017), we determined the temperature stability at 30°C in the time course of 100 min at 50 mM glucose and 5 mM ATP. Protein concentrations varied over separate preparations (30–300 µg/mL) without having an effect on the protein stability. We extended these measurements in the follow-up study by Těšínský *et al.* (2019), in which we measured thermostability of the wild type GST-GCK and its somatic cancer-associated variations at 30°C, 37°C, 42°C and 45°C in the course of a 100 min incubation at the indicated temperature. We diluted all proteins to 100 µg·ml⁻¹. We measured the GCK activity in the presence of 50 mM glucose and 5 mM ATP.

We calculated, based on the determined kinetic variables ($S_{0.5}$, n_H , k_{cat} and ATP K_M), the relative activity index (RAI) and the glucose threshold for glucose-stimulated insulin release (GSIR-T). The RAI values serve as a direct comparison of the GCK mutants with the wild-type enzyme. The equation has been previously published (Matschinsky, 2009). We employed a minimal mathematical model, which reflects the kinetic characteristics of the wild-type GCK and its mutant forms, as well as the stability coefficient and adaptation through the expression coefficient to predict the β -cell threshold for GSIR. The previously published consensual assumptions were fulfilled (Davis *et al.*, 1999; Matschinsky *et al.*, 2000).

pH optimum of GCK and influence of ATP on buffering capacity (Šimčíková & Heneberg, 2019)

To test the buffering capacity of commonly used enzyme assay buffers according to changing ATP concentrations, we prepared the reaction mixtures as follows: 1 mL of the respective buffer; 0.4 mL of the GST-GCK elution buffer, with or without the tested enzyme; 0.1 M ATP in various volumes; and dH₂O added to adjust the total volume to 2 mL. The composition of the elution buffer was as follows: 2.6 mM NADP, 0.1 mL 1 M glucose, 0.2 mL 50 mM Tris, 200 mM KCl, and 5 mM DTT; pH adjusted to 8.0. We kept all solutions at 23°C,

except for ATP and NADP, which were kept on ice. In some cases, we observed the shift in pH towards more acidic values after the addition of NADP. The amount of NADP was constant in all mixtures; therefore, any other observed changes in pH were caused only by changing ATP concentration. The ATP solution was added to the buffers in a form of a 100 mM aqueous solution that was prepared directly from the ATP disodium salt hydrate powder, without any adjustment of its pH and without the addition of any salts. ATP was always added shortly before the experiments to avoid any potential issues with its stability.

We conducted measurements at 1 mM ATP, 50 mM glucose, 100 mM Tris, for pH range of 7.5–8.8 or 100 mM glycine for pH range of 8.6–10.3. We measured HK2 and GST-GCK activity using a coupled reaction with glucose-6-phosphate dehydrogenase as described previously (*Liang et al., 1995; Davis et al., 1999*). In the case of HK2, we measured enzymatic activity in the range of 0–2 mM glucose, unlike GST-GCK, which we measured in the range of 0–150 mM glucose. We prepared all the buffers and measured the enzyme kinetics at 23°C, thereby excluding effects of temperature on pH of the solutions used.

Prediction methods used for GCK variations (*Šimčíková et al., 2017; Těšínský et al., 2019*)

For the prediction analyses, we used a protein identifier (GCK NCBI code: NP_000153.1; GCK Swiss-Prot code: P35557), or directly an amino acid sequence in FASTA format. We retrieved data related to the nonsynonymous single nucleotide variations (nonsynonymous substitutions, abbreviated as nsSNVs) in the expressed region of the GCK gene from the Ensembl (<http://www.ensembl.org/>), dbSNP (<https://www.ncbi.nlm.nih.gov/projects/SNP/index.html>), UniProtKB (<http://www.uniprot.org>) and HGMD (*Stenson et al., 2014*) databases and from a systematic review of the literature published in 2009-2017 and listed in the Web of Science database (<http://apps.webofknowledge.com>). We obtained the

structure of the closed form of GCK (Protein Data Bank (PDB) ID: 1V4S; *Kamata et al., 2004*) from PDB (<http://www.rcsb.org/pdb/home/home.do>).

We employed methods that use evolution-based sequence information (SIFT, PhD-SNP) and methods that take into account the chemical and physical characteristics of amino acids (Align-GVGD) or protein structural attributes combined with multiple sequence alignment-derived information (PolyPhen-2, SNAP2, SNPs&GO) to predict the phenotypic effect of nonsynonymous substitutions. A single amino acid substitution can result in a notable change in the protein stability, which is represented by a change in its Gibbs free energy ($\Delta\Delta G$) upon folding. Therefore, we employed two predictors that focus on the stability properties of the nonsynonymous substitutions, I-Mutant 3.0 and PoPMuSiC 2.1. We also used EVmutation to evaluate the efficiency of the epistatic approach for protein function and the stability prediction.

The Sorting Intolerant From Tolerant (SIFT) method (*Kumar et al., 2009*) is based on the hypothesis that protein evolution is correlated with protein function. Functionally relevant amino acids should be conserved in the protein family, whereas less important positions should be diverse. The SIFT Human Protein predicts whether nonsynonymous substitutions affect the protein function for all Ensembl transcripts with an assigned ENSP number (GCK ENSP: ENSP00000384247). Based on their scores, the substitutions are considered to be damaging (≤ 0.05) or tolerated (> 0.05), ideally with median sequence information (also referred as the median conservation value) between 2.75 and 3.25. The median sequence information provides an assessment of the confidence, and SIFT computes the conservation value at each position in the alignment. The conservation value ranges from 0, which means that all 20 amino acids are at that position, to 4.32, which means that the position is completely conserved. A sufficient diversity within the aligned sequences is maintained by median sequence information of ~ 3.0 .

The PolyPhen-2 (Polymorphism Phenotyping v2) method (*Adzhubei et al., 2010*) estimates the probability of the nonsynonymous substitution to adversely affect protein function based on sequence, phylogenetic and structural features. The nonsynonymous substitution is predicted as probably damaging (0.85–1.00), possibly damaging (0.15–0.84) or benign (<0.15). We identified the nonsynonymous substitution effect according to the HumDiv score. The model was trained on a dataset that involved known effects of damaging alleles that cause human Mendelian diseases that are annotated in the UniProtKB database.

SNAP2 (Screening for non-acceptable polymorphisms) (*Hecht et al., 2015*) is a neural network-based classifier that predicts the impact of nonsynonymous substitutions based on evolutionary information, structural features and solvent accessibility. The score ranges from -100 (strong neutral prediction) to +100 (strong effect prediction).

PhD-SNP (Predictor of human Deleterious Single Nucleotide Polymorphisms) (*Capriotti et al., 2006*) is a support vector machine (SVM)-based classifier that distinguishes disease-related nonsynonymous substitutions from neutral ones by reflecting the nature of the substitution and properties of the neighboring sequence environment. The method was optimized using a dataset of neutral and deleterious variations taken from the UniProtKB/Swiss-Prot.

The SNPs&GO method (*Calabrese et al., 2009*) is based on a principle very similar to PhD-SNP. In contrast to PhD-SNP, the SNPs&GO also takes into account protein function information that is defined by Gene Ontology (GO) (*Ashburner et al., 2000; The Gene Ontology Consortium, 2015*) terms. GO terms are directly retrieved only if a Swiss-Prot code is used. If GO terms are not included and only protein sequence input is available, the accuracy of the method is thought to be lower and comparable with PhD-SNP (*Calabrese et al., 2009*).

Align-GVGD (*Tavtigian et al., 2005; Mathe et al., 2006*) classifies the amino acid substitutions and their functional effect according to the “C-score” that ranges from 0 (neutral)

to C65 (deleterious). The C-score is based on the cross-species protein multiple sequence alignment with a comparison of the physical and chemical characteristics of amino acids. The Align-GVGD combines the GV (Grantham variation) and GD (Grantham deviation) score. We expressed the evolutionary conservation of the amino acid sequence of the pancreatic isoform of GCK in the form of a GV score, which was based on the alignment of the human GCK protein sequence with the GCK sequences of 12 other vertebrate species. Because the GCK sequence is highly conserved, the alignment included not only mammals (three species) but also birds, amphibians, reptiles and fish. To calculate the GV score, we used the multiple sequence alignment, which was formed using ClustalW in MEGA6, and built on the following sequences: *Homo sapiens* NP_000153.1; *Mus musculus* NP_034422.2; *Rattus norvegicus* XP_006251241; *Bos taurus* NP_001095772.1; *Danio rerio* NP_001038850; *Cyprinus carpio* ACD37722; *Meleagris gallopavo* XP_010725006; *Aquila chrysaetos* XP_011573674; *Ficedula albicollis* XP_005057963; *Xenopus laevis* NP_001079298; *Nanorana parkeri* XP_018422966; *Anolis carolinensis* XP_003224263; *Lepisosteus oculatus* XP_006625388. Positions with zero GV score have the same amino acids across all species and are thus invariant, whereas the GV increases when the alignment demonstrates evidence for variation in the particular residue. The *GCK* gene does not have any insertions or deletions of amino acids in the studied species, except for the N- and C-terminal parts of the molecule; thus, nearly all the variability was assigned to nonsynonymous substitutions.

I-Mutant 3.0 (Capriotti *et al.*, 2008) was designed to estimate the protein stability change caused by nonsynonymous substitutions. The tool was trained on a dataset built on the information from ProTherm (Kumar *et al.*, 2006), which is a comprehensive thermodynamic database of experimental data for wild-type and mutant proteins. Based on the protein structure or the sequence, the difference ($\Delta\Delta G$ value) between the unfolding Gibbs free energies of the mutated and wild-type protein is calculated. In the present study, we based the $\Delta\Delta G$ values on

the protein structure of GCK (PDB ID: 1V4S; *Kamata et al., 2004*). Nonsynonymous substitution with $\Delta\Delta G > 0.5$ kcal mol⁻¹ are considered to be largely stabilizing, and those with $\Delta\Delta G < -0.5$ kcal mol⁻¹ are predicted as largely destabilizing. Other nonsynonymous substitution with $\Delta\Delta G$ in the range from -0.5 to 0.5 kcal mol⁻¹ have a weak effect (*Capriotti et al., 2008*).

Another web server that allows predicting the thermodynamic stability changes upon the nonsynonymous substitution is PoPMuSiC-2.1 (*Dehouck et al., 2011*). This method reflects the solvent accessibility of the mutated residue. The predictions are derived from the structure of the target protein (GCK PDB ID: 1V4S). The $\Delta\Delta G$ values lower than 0 kcal mol⁻¹ are assigned to stabilizing nonsynonymous substitutions, and those that are higher than 0 kcal mol⁻¹ are assigned to destabilizing nonsynonymous substitutions.

The prediction method EVmutation (*Hopf et al., 2017*) exploits the epistatic approach. Thus, it takes into account explicitly modelling of interactions between all the pairs of residues in the proteins and bases in RNAs to predict nonsynonymous substitution effects. Within validation, EVmutation predictions were compared with outcomes from 34 high-throughput mutagenesis experiments. The EVmutation scores (ΔE) below 0 are assigned to deleterious nonsynonymous substitutions, values above 0 correspond to beneficial nonsynonymous substitutions, and values equal to 0 correspond to neutral nonsynonymous substitutions.

The developers of all prediction methods suggested interpreting the resulting predictions using arbitrary scores as threshold values. We presented the calculations using these arbitrarily suggested interpretations and thresholds in Table 2. However, arbitrary thresholds were associated with extreme uncertainty and overestimated the effects of neutral nonsynonymous substitutions. Nevertheless, we found that three prediction methods, PolyPhen-2, SNAP2 and EVmutation, allowed differentiating at least in part between the neutral and MODY-associated nonsynonymous substitutions when considering their numerical outcomes. Thus, for these three methods, we computed (PolyPhen-2 and SNAP2) or retrieved (EVmutation) predictions for all

possible amino acid exchanges within the GCK molecule, irrespectively on whether they are already known from humans or not. For SNAP2, we retrieved 8,837 predictions with mean value 4.54 ± 0.63 (min -99, max 96, median 13, 25th percentile -52, 75th percentile 58). For PoPMuSiC 2.1, we retrieved 8,856 predictions with mean value 1.10 ± 0.01 (min -1.88, max 5.77, median 0.85, 25th percentile 0.32, 75th percentile 1.70). For EVmutation, we retrieved 8,191 predictions with mean value -5.35 ± 0.03 (min -10.15, max 4.10, median -5.36, 25th percentile -7.06, 75th percentile -3.79). We applied two types of adjusted thresholds in order to be able to predict nonsynonymous substitutions, which are likely to serve as causative MODY nonsynonymous substitutions, and which are likely to be benign or activating. We calculated the thresholds by computing the medians and SDs of scores for nonsynonymous substitutions, which do not cause any monogenically inherited disease. We calculated the threshold for predicting the MODY-associated nonsynonymous substitutions as median of scores for nonsynonymous substitutions, which do not cause any monogenically inherited disease, with the addition of 2 SDs. We calculated the threshold for predicting the benign (or activating) nonsynonymous substitutions as the median value of scores for nonsynonymous substitutions, which do not cause any monogenically inherited disease. We used these evidence-based thresholds for a further validation of these methods.

Databases of missense variations in genes encoding proteins associated with Mendelian diseases (*Šimčíková & Heneberg, subm.*)

We assembled two curated databases of missense variations in genes encoding proteins associated with Mendelian diseases to establish and validate the model. When establishing the model, we recognized three categories of variations: 1) “DAVs” represented variations with available evidence of an association with Mendelian diseases. 2) “Partial phenotype-associated” variations were reported to be associated with partial (incompletely manifesting) phenotypes of

the same Mendelian diseases. And 3) “No phenotype-associated” variations (NPAVs) were variations with conclusive evidence of the absence of any clinical phenotype associated with their carriers.

In addition to the clinically observed variations, we calculated and analyzed the predictions for theoretical variations, i.e., variations that have not been clinically observed. We sorted the variations according to a) their localization within/outside protein domains, b) the presence and class of enzymatic activity of the protein, c) the number of nucleotide changes needed to obtain the variation of interest, and d) the American College of Medical Genetics and Genomics (ACMG) classification criteria (Nykamp *et al.*, 2017).

We selected genes encoding proteins associated with Mendelian diseases according to the availability of a protein structure, inheritance of diseases, and sufficient numbers of clinically observed missense variations (at least nine missense DAVs and at least six missense NPAVs in a region for which the protein structure was available). We retrieved data from the Online Mendelian Inheritance in Man (OMIM; <https://omim.org/>), UniProtKB/Swiss-Prot (<http://www.uniprot.org/>), Protein Data Bank (PDB; <https://www.rcsb.org/>) and Human Gene Mutation Database (HGMD; <http://www.hgmd.cf.ac.uk>). We obtained the evidence for the presence of NPAVs from the ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) and Ensembl (<http://www.ensembl.org/>) databases. We completed information with frequencies of variations and protein domains obtained from the Exome Aggregation Consortium browser (ExAC; <http://exac.broadinstitute.org/>) and the Pfam (<http://pfam.xfam.org/>) database, respectively.

We verified all ambiguous data in the primary literature sources. If we observed conflicting evidence or if conclusive evidence was not available, we removed the variations from the analyses. The factors that led to the removal of variations from the analyzed datasets are listed below. 1) The evidence for only non-Mendelian diseases (e.g., Parkinson disease) was manifested in the carriers of the variation. 2) The variations were listed as benign or likely

benign in ClinVar, with high frequencies ($f > 8$) in ExAC, and thus were classified as 1B or higher according to the ACMG criteria for high-quality and abundant data (*Richards et al., 2015*). 3) The variations were listed as “DM?” in the HGMD database. These variations denote “a probable/possible pathological mutation, reported to be pathogenic in the corresponding report, but for which (1) the author has indicated that there may be some degree of uncertainty; (2) the HGMD curators believe greater interpretational caution is warranted; or (3) subsequent evidence has appeared in the literature which has called the putatively deleterious nature of the variant into question” (*Stenson et al., 2014*). 4) Variations for which a disagreement occurred between HGMD (classified as “DM”) and ClinVar (classified as “benign” or “likely benign”).

We used the key provided in Table 1 to assign of the clinically observed variations. We selected all clinically observed variations, which we used to set the thresholds, using the key described above. Additionally, we included the GCK variations resulting from the systematic literature review (*Šimčíková et al., 2017*). We classified nine variations as NPAVs based on the recent literature (*Liu et al., 2009; Steele et al., 2011; Chellapa et al., 2012; Houlleberghs et al., 2016; Maxwell et al., 2016; Walsh et al., 2016*). We included the hemoglobin variations, which were classified as likely non-phenotypic in the HGMD database, in the NPAVs.

Table 1. The key used to assign of the clinically observed variations. Abbreviations used: DIS – disease-associated; PART – partial phenotype-associated; NO PHEN – no phenotype-associated; EXCL – excluded ambiguous data.

1a)	In HGMD, the variation is absent.	2
1b)	In HGMD, the variation is present, but causes “no phenotype” according to dbSNP.	NO PHEN
1c)	In HGMD, the variation is present and is defined as a “disease causing mutation”.	4
1d)	In HGMD, the variation is present but has with definitions other than those listed in 1b) and 1c)	2

2a)	In ClinVar, the variation is present and defined as “benign”, “likely benign” or “variants of uncertain significance” (VUSs).	NO PHEN
2b)	In ClinVar, the variation is absent or present, with definitions other than those listed in 2a).	3
3a)	In Ensembl, the variation is present but has no associated phenotype.	NO PHEN
3b)	In Ensembl, the variation is present and associated with a phenotype.	5
4a)	In ClinVar, the variation is present and defined as “benign” or “likely benign”.	EXCL
4b)	In ClinVar, the variation is present but not defined as “benign” or “likely benign”.	5
5a)	In HGMD, all variations classified as “disease-causing mutations” within the respective gene are associated with a single disease or syndrome with a Mendelian inheritance pattern.	DIS
5b)	In HGMD, the variations classified as “disease-causing mutations” within the respective gene are associated with two diseases with a Mendelian inheritance pattern, one caused by the activating and the other by inactivating variations (e.g., erythrocytosis vs anemia).	DIS
5c)	In HGMD, the variations classified as “disease-causing mutations” within the respective gene are associated with two diseases with a Mendelian inheritance pattern, both of which are caused by variations exerting similar effects with a different intensity (e.g., Menkes syndrome vs occipital horn syndrome or Duchenne vs Becker muscular dystrophy); variations cause a complete phenotype.	DIS
5d)	In HGMD, the variations classified as “disease-causing mutations” within the respective gene are associated with two diseases with a Mendelian inheritance pattern, both of which are caused by variations exerting similar effects with a different intensity (e.g., Menkes syndrome vs occipital horn syndrome or Duchenne vs Becker muscular dystrophy); variations cause the less pathological phenotype.	PART

The variations classified in ClinVar as VUS (n = 404) were subjected to the analysis using EVmutation, and SNAP2 scores shifted slightly but significantly towards their pathogenicity compared to the variations classified as benign or likely benign (n = 1589): EVmutation mean \pm SD -4.21 ± 2.54 vs -3.84 ± 2.38 , *t*-test *p* = 0.003; SNAP2 mean \pm SD -0.9 ± 57.34 vs -12.57 ± 55.85 , *t*-test *p* < 0.001. Based on these calculations, we excluded the

hemoglobin variations that were classified as likely non-phenotypic in HGMD ($n = 100$). These variations received EVmutation scores, but not SNAP2 scores, similar to VUS (EVmutation mean \pm SD -4.26 ± 2.25 , t -test vs VUS $p > 0.05$; SNAP2 mean \pm SD 15.58 ± 42.69 , t -test vs VUS $p = 0.003$).

All variations included in the dataset we used to establish the model were classified according to the ACMG criteria (Nykamp *et al.*, 2017), differentiating between those classified as benign (1B, 3B and 5B) and pathogenic (0.5P and 1P).

We retrieved clinical information on 7178 missense variations located within the coding sequences of 44 genes that, if mutated, cause Mendelian diseases. We included the following genes in the dataset we used to validate the model: *AR*, *ATP7A*, *BMPR2*, *BTK*, *CD40LG*, *CDKL5*, *CPOX*, *CYBB*, *DCX*, *DMD*, *EDA*, *ELANE*, *F9*, *FHL1*, *FLNA*, *G6PD*, *GCK*, *GCH1*, *GLA*, *HBB*, *HDAC8*, *HMBS*, *HNF4A*, *HPRT1*, *HSPB1*, *IDS*, *IL2RG*, *ITGA2B*, *KIT*, *MECP2*, *MSH2*, *OTC*, *PDHA1*, *PROC*, *PTEN*, *PTPN11*, *RET*, *SERPING1*, *SH2D1A*, *STK11*, *TGFBR2*, *TP63*, *TTR* and *UROD*. We limited all the analyzed missense variations to those parts of the genes for which structural information was available. We designated 4546 variations as “DAVs”, because the evidence for their associations with Mendelian diseases was available. We designated another 291 variations as “partial phenotype-associated”, because the evidence for their association with partial (incompletely manifesting) phenotypes of the same Mendelian diseases was available. We designated 2093 missense variations as “NPAVs”, because conclusive evidence of the absence of any clinical phenotype associated with their carriers was available. We removed 248 (3.5%) missense variations from the analyses due to inconsistent, insufficient or anomalous data on the phenotypes reportedly associated with these variations. Data reliability in databases appears to be a challenge to the construction of the dataset. Standardized forms of annotations do not currently exist. Additionally, submission processes differ among the databases, ranging from individual to bulk submissions, and are rarely checked

for consistency with previously published peer-reviewed studies (*Maxwell et al., 2016*). Therefore, the construction of the comprehensive dataset also prevented or considerably decreased the risk of biases that might arise from errors of omission and commission in databases.

Selection of genes to validate the model (*Šimčíková & Heneberg, subm.*)

We established the validation dataset consisting of 1723 variations in 63 additional genes associated with autosomal dominant or autosomal recessive diseases to validate the newly reported approach on an independent set of proteins that are associated with Mendelian diseases. These 63 genes were not included in the dataset that was used to establish the model. We populated the dataset based on the classifications of variations retrieved from ClinVar. We also verified the allele counts in the ExAC browser, but this information was only available for a limited number of variations in this dataset. Thus, this information was not used in the analyses. The genes included in the dataset that was used to validate the model were: *AARS, ABCC6, ALDH18A1, ARSB, AVP, CASR, CFTR, CLCN1, CLCN7, COL7A1, DNM2, DSP, DYNC1H1, ELOVL4, FBN1, FGF23, FGFR3, GALNS, GBA, GJB2, GJA3, GLB1, GNE, GUCY2D, GUSB, HEXA, HGSNAT, IMPDH1, KCNA1, LMNA, LMNB1, LRP5, MARS, MPZ, MYH14, MYH3, MYH7, MYH9, MYO6, NAGLU, NOTCH3, NR3C2, OPA1, PGFRB, PKD1, PKD2, POLG2, PRKCG, PRPF8, RAF1, RYR1, SGSH, SLC4A1, SMPD1, SOS1, SOS2, SPAST, STAT1, STAT3,TECTA, TERT, VCP* and *YARS*. The dataset was composed of the following numbers of variations: 33 benign, 53 benign / likely benign variations, 58 likely benign variations, 475 likely pathogenic variations, 104 likely pathogenic / pathogenic variations and 1000 pathogenic variations.

Prediction methods – extended analysis (*Šimčíková & Heneberg, subm.*)

We used the pre-computed predictions from EVmutation that were listed according to the UniProtKB/Swiss-Prot accession numbers. We computed the predicted effects of amino acid changes identified using SNAP2 according to the NCBI code belonging to relevant protein isoforms. We selected the protein structures with a resolution lower than 2.7 Å (except GCH1 and PROC) and used their PDB codes in the prediction computations employing PoPMuSiC 2.1. In addition to the clinically confirmed variations, we calculated and analyzed the predictions for theoretical variations, i.e., variations that were not clinically observed. We performed these calculations for the protein regions identical to those, we used to analyze the clinically observed variations. We sorted the variations according to a) their localization within/outside of protein domains, b) the presence and class of enzymatic activity of the protein, and c) the number of nucleotide changes needed to obtain the variation of interest. When sorting the variations according to the latter criterion, we split theoretical variations into impossible (157,639 variations) and possible variations (63,698 variations) according to the method reported by *Bromberg et al. (2013)*. They defined “impossible” amino acid variations as those that require a change of two or three nucleotides in the codon, whereas “possible” variations were defined as amino acids variations that require a change in only a single nucleotide.

GV approach – extended analysis (*Šimčíková & Heneberg, subm.*)

We assembled the MSAs by implementing the paradigm associated with variants of uncertain significance (VUS), which claims that the variations are considered VUSs if an amino acid residue that is conserved in the corresponding protein in other mammals is altered (*Richards et al., 2015*). Thus, for each analyzed protein, we prepared the MSA that contained amino acid sequences of ten mammalian orthologs of the respective gene. Typically, we included a dominant human isoform of the respective protein and complemented it with the

corresponding isoform reported from two species of primates (*Primates*) and one sequence each from carnivores (*Carnivora*), bats (*Chiroptera*), rodents (*Rodentia*), even-toed ungulates or cetaceans (*Cetartiodactyla*) and insectivorous mammals (*Eulipotyphla*, which is still listed as *Insectivora* in the NCBI Nucleotide database). The remaining two orthologs were both represented by marsupials (*Metatheria*) or by one marsupial and one monotreme (*Monotremata*), avoiding monotreme sequences when high-quality reads were not available in the NCBI GenBank database.

Additionally, we tested two representative genes, *AR* and *PTEN*, to determine whether the addition of more evolutionarily distant sequences and the resulting increase in variability led to an improved correspondence of GV scores with disease associations of analyzed variations. We used the maximum likelihood method to estimate evolutionary divergence in amino acid sequences predicted to be encoded by *AR* and *PTEN* among selected taxonomic groups. For *AR*, we tested 29 amino acid sequences of *AR* orthologs, including the orthologs from ten mammalian species, as specified above. The more evolutionarily distant orthologs included sequences from *Testudines* (three species), *Amphibia* (three species), *Crocodylia* (two species), *Squamata* (four species), *Aves* (three species), *Euteleostomi* (three species) and *Chondrichthyes* (one species). The NCBI Blast search did not retrieve orthologs that would be homologous with *AR* from more evolutionarily distant species. The *PTEN* protein is more evolutionarily conserved, which allowed us to include more distant taxa. The resulting dataset comprised 31 orthologs, ten of which were from the mammalian species listed above, and others consisted of orthologs from the following taxa: *Aves* (three species), *Squamata* (three species), *Archelosauria* (three species), *Teleostei* (three species), *Chondrichthyes*, *Coelacanthiformes*, *Amphibia*, *Brachipoda*, *Gastropoda*, *Mollusca*, *Echinozoa*, *Arachnida* and *Insecta* (one species each). We aligned the amino acid sequences using ClustalW (gap opening penalty of 5 and gap extension penalty of 0.1 for pairwise alignments, gap extension penalty of 0.2 for multiple

alignments, and gap separation distance of 4). We manually corrected the alignments for any inconsistencies and replaced shorter sequences with more appropriate sequences. We used only sequences of identical lengths for further analyses. We used the resulting MSAs to calculate the GV scores. For the AR and PTEN alignments, we performed maximum likelihood fits of the 48 amino acid substitution models, excluding positions containing gaps. For each model, we calculated the Bayesian information criterion, corrected Akaike information criterion and maximum likelihood values. For AR, we analyzed 29 sequences with 380 positions in the final dataset. For PTEN, we analyzed 31 sequences with 342 positions in the final dataset. We used best-fit models for the subsequent phylogenetic analyses and evolutionary divergence calculations. When building the trees, we constructed the initial tree using a neighbor-joining algorithm. We built the trees based on both AR and PTEN sequences using the Jones-Taylor-Thornton model. We modeled the non-uniformity of evolutionary rates among sites using a discrete Gamma distribution (+G) with five rate categories. We applied a bootstrapping procedure with 1,000 replicates. We used the maximum likelihood method to estimate evolutionary divergence in the amino acid sequences of AR and PTEN orthologs among selected taxonomic groups. We calculated the number of base differences per site by averaging all sequence pairs between groups (distance) \pm SE and employed a bootstrapping procedure with 1,000 replicates. The models used to estimate inter- and intrasite evolutionary divergence were identical to the models used to construct the respective trees.

Prediction method REVEL (*Šimčíková & Heneberg, subm.*)

We calculated the sensitivity and specificity of the predictions retrieved from REVEL to test whether the issue of low specificity is associated with the outcomes of individual computational algorithms or whether it also affects the data obtained using state-of-the-art consensus classifiers (*Ioannidis et al., 2016*). We used REVEL to test a subset of 21 genes from

the dataset that was used to establish the model: *GCK*, *AR*, *PTEN*, *CYBB*, *HNFA4A*, *HBB*, *MECP2*, *HDAC8*, *RET*, *PTPN11*, *HPRT1*, *CD40LG*, *CDKL5*, *CPOX*, *DCX*, *DMD*, *EDA*, *UROD*, *TTR*, *FLNA* and *HSPB1*. We provided REVEL scores for 2721 variations, of which 1570 were DAVs, 241 manifested partial phenotypes, and 910 were NPAVs. For the aforementioned genes, we tested the identical set of variations as used to establish the model, except for PTEN p.P103Q, PTEN p.A137F, and four GCK variations, representing amino acid substitutions caused by substitutions of two or three nucleotides. We obtained the REVEL scores from the pre-computed database of REVEL scores that are available for all missense variations retrieved from dbNSFP v2.7, as provided by the authors of REVEL (Ioannidis *et al.*, 2016).

Statistical methods – GCK and prediction methods (Šimčíková *et al.*, 2017)

We analyzed the data by one-way ANOVA with Tukey's post-tests and computed the $S_{0.5}$, Hill coefficient n_H , k_{cat} and ATP K_M via non-linear regression analyses. We obtained IC_{50} using four parameters logistic curve fitting. Multiparametric analyses included the detrended correspondence analyses. We calculated Pearson product moment correlation coefficients and Spearman rank order correlation coefficients in order to correlate the numerical outputs of EVmutation, PoPMuSiC 2.1 and SNAP2 prediction methods applied to total hypothetical GCK nonsynonymous substitutions for which the outcomes of all the three prediction methods were available ($n_{PoPMuSiC\ 2.1 / EVmutation} = 8,493$ nonsynonymous substitutions, $n_{SNAP2 / PoPMuSiC\ 2.1}$ and $SNAP2 / EVmutation = 8,189$ nonsynonymous substitutions). We also calculated the two correlation coefficients in order to compare GV with the frequency of families ($n = 465$ residues, of that 279 residues were disease-associated (1596 disease-associated families) and 164 residues were not evolutionarily conserved). Data were shown as means \pm SE, unless stated otherwise. We performed the calculations and plotted the figures in PAST 2.14 and SigmaPlot 12.0.

Statistical methods – GCK (*Těšínský et al., 2019; Šimčíková & Heneberg, 2019*)

The results were shown as the mean \pm SEM. Following a Shapiro-Wilk normality test and Levene's equal variance test, data were either analyzed using one-way ANOVA or Kruskal-Wallis ANOVA on ranks. For the post-tests, we used Dunnett's multiple comparison tests. We calculated the Pearson product moment correlation coefficient and the Spearman rank order correlation coefficient in order to correlate the IC₅₀ of GlcNAc and the Hill coefficient.

Statistical analyses – extended prediction analysis (*Šimčíková & Heneberg, subm.*)

We calculated the evidence-based thresholds as medians \pm 2 \times SD, which should encompass approximately 95% of the pool of variations used to calculate the threshold. We calculated two types of these thresholds. The sensitivity threshold (true positive rate) was calculated based on the 95% chance of confirming the association of a tested theoretical variation with the respective disease based on the distribution of prediction scores for known DAVs. The specificity threshold (true negative rate) was calculated based on the 95% chance of confirming the absence of an association of a tested theoretical variation with the respective disease based on the distribution of prediction scores for known NPAVs.

We calculated the weighted means of the scores resulting from the tested prediction methods by assigning each predictor a weight ranging from -100 to +100, where 0 was a threshold and 100 was the maximum value observed within the respective dataset (EVmutation range -12.933 – 3.8104, SNAP2 range -98 – 99, and PoPMuSiC 2.1 range -1.90 – 5.64), and by averaging the values obtained from each of the prediction methods.

We tested the differences between predictions between DAVs and NPAVs, and for domain-associated and other amino acids using a one-tailed *t*-test. Differences in the numbers of DAVs and NPAVs in individual domains were determined using one-tailed *t*-tests with

Bonferroni's correction. We tested the differences between variations associated with particular classes of enzymes and proteins without enzymatic functions, and between categories of possible and impossible theoretical variations using the Kruskal-Wallis one-way ANOVA on ranks with Dunn's post-tests (the Kolmogorov-Smirnov normality test yielded $p > 0.05$ for each comparison). We analyzed the difference in the frequency of DAVs and NPAVs among possible and impossible theoretical variations using the χ^2 test, with the number of possible variations normalized to the number of impossible variations. We assessed the differences between DAVs (including multiple phenotypes alone), partial phenotype-associated and NPAVs using the Kolmogorov-Smirnov normality test followed by one-way ANOVA with Tukey's post-tests or Kruskal-Wallis one-way ANOVA on ranks with Dunn's post-tests when the normality tests failed. We did not evaluate phenotypes with less than five associated variations. The data are shown as means \pm SD, unless indicated otherwise. We performed all calculations using SigmaPlot 12.0, and conducted phylogenetic analyses using MEGA 5.2.

Preparation of hexokinase knockout cell lines using CRISPR/Cas9

For CRISPR/Cas9, we used the plasmid pSpCas9(BB)-2A-GFP (*Ran et al., 2013a*). This plasmid encodes sgRNA, Cas9 and GFP. We designed sgRNAs targeting into *HK1* and *HK2* genes using the CHOPCHOP tool (*Labun et al., 2016, 2019*) (Table 2).

Table 2. Newly designed sgRNAs that target into *HK1* and *HK2* genes.

Name of sgRNA	Targeted gene (exon)	5'→3' sequence encoding sgRNA
sgRNA14	<i>HK1</i> (exon 1)	CTGCGCGGCGATCATGCTGG
sgRNA16	<i>HK1</i> (exon 3)	GAGAACATCGTGCACGGCAG
sgRNA22	<i>HK1</i> (exon 3)	TTGCACCCGCAGAATTCGAA
sgRNA3	<i>HK2</i> (exon 7)	GATGCGCCACATCGACATGG
sgRNA17	<i>HK2</i> (exon 2)	ACCGCTTAGAGATCTCCAAG
sgRNA30	<i>HK2</i> (exon 5)	CGTTGTGGCTCTGATCCGGA

We annealed and cloned the sequences encoding sgRNAs and their complementary oligonucleotides on the Golden-Gate sgRNA cloning protocol (https://media.addgene.org/cms/filer_public/3e/e1/3ee1ce9c-99f9-4074-9a28-109d34971471/zhang-lab-sam-cloning-protocol.pdf).

We have grown the adherent ovarian cancer cell line TOV-112D in a mixture of MCDB 105 medium and Medium 199 (1:1, v/v) containing a final concentration of 15% fetal bovine serum (FBS). We have grown the adherent human embryonic kidney cell line HEK293T in DMEM medium containing 10% FBS. A day before transfection, we seeded the cells into the 6-well plate (3×10^5 cells/well). We transfected the cells using either Lipofectamine 2000 (ThermoFisher Scientific, Waltham, MA) or poly(ethylenimine) (PEI, MW 25000;

Polysciences, Hirschberg an der Bergstraße, Germany) solution according to the manufacturers' instructions. A day after transfection, we performed a single-cell sorting according to GFP expression using FACS BD Aria. We kept the sorted cells in culture until we had enough cells of each clone for cryopreservation, DNA isolation and preparation of lysates for Western blotting.

Restriction analysis of CRISPR/Cas9 clones

We designed the primers for PCR using the CHOPCHOP v3 tool (Labun *et al.*, 2016, 2019). The PCR product contained the respective targeted site for Cas9 and restriction site for the respective restriction endonuclease that detected changes in the targeted site because of distinct cleavage of the wild-type and inaccurately repaired site (Table 3).

Table 3. PCR primers and restriction enzymes used for restriction analyses.

Name of sgRNA	Targeted gene (exon)	5'→3' Forward primer	5'→3' Reverse primer	Restriction enzyme
sgRNA14	<i>HK1</i> (exon 1)	GGAGGAGGAGGAGGAGGAG	GGCTCACCTTTTTGACCTGG	NA
sgRNA16	<i>HK1</i> (exon 3)	TATGTGGCTTCCCCTTAACATT	TCTATGAGGGACTCTTTCCA GC	AleI
sgRNA22	<i>HK1</i> (exon 3)	TATGTGGCTTCCCCTTAACATT	TCTATGAGGGACTCTTTCCA GC	EcoRI
sgRNA3	<i>HK2</i> (exon 7)	GTATAAGAGGGAAGAGGGGT GG	CATGTCAATCTCCTGGTCAA AC	HpyAV
sgRNA17	<i>HK2</i> (exon 2)	TCTTCCTCCTTTTTCAGGTTGA	GAGCAAAGCCAACTAAATCA CC	BstYI
sgRNA30	<i>HK2</i> (exon 5)	TTCCAGAGTTTCTGGTCTCAT	TTAAGCTCCACGTAAGCAAA CA	BspEI

NA = not applicable.

We isolated the DNA using QuickExtract DNA Extraction Solution (Lucigen, Middleton, WI). We performed PCR reactions using Herculase II Fusion DNA polymerase (Agilent, Santa Clara, CA). We used the restriction enzymes according to the New England Biolabs instructions.

Western blotting and immunodetection

The TOV-112D and HEK293T cells were trypsinized or not, respectively, centrifuged, lysed in 1× SDS-PAGE loading buffer and incubated at 99°C for 15 min. We used the lysates for SDS-PAGE (*Green & Sambrook, 2012*). We transferred the SDS-PAGE gels on a nitrocellulose membrane in Towbin buffer (25 mM Tris, 192 mM glycine, pH 8.3, 20% methanol) at 100 V and 4°C for an hour.

Afterwards, we blocked the nitrocellulose membrane in 5% milk, PBS, 0.05% Tween 20 for an hour. Then, we incubated the membrane in a primary antibody diluted in 5% milk or 5% BSA, PBS, 0.05% Tween 20 overnight at 4°C. Subsequently, we washed the membrane 3-times with PBS, 0.05% Tween 20 and incubated with a secondary antibody conjugated with horseradish peroxidase (HRP) for 45 min at room temperature. After six washing steps with PBS, 0.05% Tween 20, we incubated the membrane with a chemiluminescent substrate for HRP and performed protein detection by ChemiDoc Imaging System (Bio-Rad, Hercules, CA).

To test the CRISPR/Cas9 clones for HK1 and HK2 expression by Western blotting, we used the following primary antibodies: rabbit anti-HK1 mAb (C35C4; Cell Signaling, Danvers, MA), rabbit anti-HK2 mAb (C64G5; Cell Signaling, Danvers, MA), mouse anti-β-actin (sc-47778; Santa Cruz Biotechnology, Dallas, TX) and mouse anti-GAPDH (sc-47724; Santa Cruz Biotechnology, Dallas, TX) as loading controls. Goat anti-rabbit and anti-mouse HRP-IgG Abs were used as secondary antibodies (A6154 and A8924; Sigma-Aldrich, St. Louis, MO).

To map the changes in expression of metabolic enzymes and associated signaling pathways, we used the TOV-112D CRISPR/Cas9 HK1 KO clone coded as **E9-14-3**, and the TOV-112D CRISPR/Cas9 HK1⁺ clone coded as **D11-14-5**. The clone E9-14-3 was a result of repair induced by the action of sgRNA14-guided Cas9. The clone D11-14-5 was a control clone, in which the action of sgRNA14-guided Cas9 did not disrupt the targeted site, thus this clone was still expressing HK1. We seeded 3 x 10⁵ cells of each clone into 2 mL of medium per one

well of a 6-well plate. We used DMEM containing 1 g/L or 4.5 g/L glucose (D5523 and D7777; Sigma-Aldrich, St. Louis, MO), and the constant concentration of glutamine (4 mM). We cultivated the cells at 37°C, 5% CO₂ for three days. Then, we trypsinized the cells, lysed them and used the lysates for Western blotting as described above.

For the investigation of glycolytic enzymes, we used the following primary antibodies: rabbit anti-HK1 mAb (C35C4; Cell Signaling, Danvers, MA), rabbit anti-HK2 mAb (C64G5; Cell Signaling, Danvers, MA), rabbit anti-PFKP mAb (D4B2; Cell Signaling, Danvers, MA), rabbit anti-PGAM-1 mAb (NBP1-49532; Novusbio, Centennial, CO), rabbit PKM2 mAb (D78A4; Cell Signaling, Danvers, MA) and rabbit anti-LDHA mAb (C4B5; Cell Signaling, Danvers, MA). For detection of proteins involved in the electron transport chain, we used the following primary antibodies: rabbit anti-MTCO2 (IV) mAb (ab79393; Abcam, Cambridge, MA) and mouse Total OXPHOS human WB Ab cocktail (ab110411; Abcam, Cambridge, MA). Rabbit anti-vinculin mAb (E1E9V; Cell Signaling, Danvers, MA) and mouse anti- β -actin (sc-47778; Santa Cruz Biotechnology, Dallas, TX) were used as loading controls. Goat anti-rabbit and anti-mouse HRP-IgG Abs were used as secondary antibodies (A6154 and A8924; Sigma-Aldrich, St. Louis, MO).

For the investigation of selected carcinogenesis-associated signaling pathways we used the following primary antibodies: rabbit anti-phospho-Rictor (Thr1135) mAb (D30A3; Cell Signaling, Danvers, MA), rabbit anti-Rictor mAb (53A2; Cell Signaling, Danvers, MA), rabbit anti-phospho-Akt (Ser473) mAb (736E11; Cell Signaling, Danvers, MA), mouse anti-Akt mAb (40D4; Cell Signaling, Danvers, MA), rabbit anti-phospho-AMPK α 1 (Thr183) + anti-phospho-AMPK α 2 (Thr172) mAb (ab133448; Abcam, Cambridge, MA), rabbit anti-AMPK α 1 mAb (ab32047; Abcam, Cambridge, MA), rabbit phospho-Raptor (Ser792) mAb (2083; Cell Signaling, Danvers, MA), rabbit anti-Raptor mAb (24C12; Cell Signaling, Danvers, MA), rabbit anti-phospho-p70 S6 kinase (Thr421/Ser424) mAb (9204; Cell Signaling, Danvers, MA),

rabbit anti-p70 S6 kinase mAb (49D7; Cell Signaling, Danvers, MA), rabbit anti-phospho-S6 ribosomal protein (Ser235/Ser236) mAb (2211; Cell Signaling, Danvers, MA), rabbit anti-S6 ribosomal protein mAb (2217; Cell Signaling, Danvers, MA), rabbit anti-phospho-4E-BP1 (Ser65) mAb (9451; Cell Signaling, Danvers, MA), rabbit anti-phospho-4E-BP1 (Thr70) mAb (9455; Cell Signaling, Danvers, MA), rabbit anti-phospho-4E-BP1 mAb (Thr37/Thr46) (2855; Cell Signaling, Danvers, MA), rabbit anti-non-phospho-4E-BP1 (Thr46) mAb (89D12; Cell Signaling, Danvers, MA), rabbit anti-4E-BP1 mAb (53H11; Cell Signaling, Danvers, MA) and rabbit anti-*c*-Myc mAb (D84C12; Cell Signaling, Danvers, MA). Rabbit anti-vinculin mAb (E1E9V; Cell Signaling, Danvers, MA) and mouse anti- β -actin (sc-47778; Santa Cruz Biotechnology, Dallas, TX) were used as loading controls. Goat anti-rabbit and anti-mouse HRP-IgG Abs were used as secondary antibodies (A6154 and A8924; Sigma-Aldrich, St. Louis, MO). We calculated intensity of bands using ImageLab (Bio-Rad, Hercules, CA).

RESULTS

Daniela Šimčíková, Lucie Kocková, Kateřina Vackářová, Miroslav Těšínský, Petr Heneberg
**Evidence-based tailoring of bioinformatics approaches to optimize methods that predict
the effects of nonsynonymous amino acid substitutions in glucokinase**

Scientific Reports (2017) 7: 9499

Abstract:

Computational methods that allow predicting the effects of nonsynonymous substitutions are an integral part of exome studies. Here, we validated and improved their specificity by performing a comprehensive bioinformatics analysis combined with experimental and clinical data on a model of glucokinase (GCK): 8835 putative variations, including 515 disease-associated variations from 1596 families with diagnoses of monogenic diabetes (*GCK-MODY*) or persistent hyperinsulinemic hypoglycemia of infancy (PHHI), and 126 variations with available or newly reported (19 variations) data on enzyme kinetics. We also proved that high frequency of disease-associated variations found in patients is closely related to their evolutionary conservation. The default set prediction methods predicted correctly the effects of only a part of the *GCK-MODY*-associated variations and completely failed to predict the normoglycemic or PHHI-associated variations. Therefore, we calculated evidence-based thresholds that improved significantly the specificity of predictions ($\leq 75\%$). The combined prediction analysis even allowed to distinguish activating from inactivating variations and identified a group of putatively highly pathogenic variations (EVmutation score < -7.5 and SNAP2 score > 70), which were surprisingly underrepresented among *MODY* patients and thus under negative selection during molecular evolution. We suggested and validated the first robust evidence-based thresholds, which allow improved, highly specific predictions of disease-associated GCK variations.

SCIENTIFIC REPORTS

OPEN Evidence-based tailoring of bioinformatics approaches to optimize methods that predict the effects of nonsynonymous amino acid substitutions in glucokinase

Received: 9 June 2017
Accepted: 28 July 2017
Published online: 25 August 2017

Daniela Šimčíková, Lucie Kocková, Kateřina Vackářová, Miroslav Těšínský & Petr Heneberg 

Computational methods that allow predicting the effects of nonsynonymous substitutions are an integral part of exome studies. Here, we validated and improved their specificity by performing a comprehensive bioinformatics analysis combined with experimental and clinical data on a model of glucokinase (GCK): 8835 putative variations, including 515 disease-associated variations from 1596 families with diagnoses of monogenic diabetes (*GCK-MODY*) or persistent hyperinsulinemic hypoglycemia of infancy (PHHI), and 126 variations with available or newly reported (19 variations) data on enzyme kinetics. We also proved that high frequency of disease-associated variations found in patients is closely related to their evolutionary conservation. The default set prediction methods predicted correctly the effects of only a part of the *GCK-MODY*-associated variations and completely failed to predict the normoglycemic or PHHI-associated variations. Therefore, we calculated evidence-based thresholds that improved significantly the specificity of predictions ($\leq 75\%$). The combined prediction analysis even allowed to distinguish activating from inactivating variations and identified a group of putatively highly pathogenic variations (EVmutation score < -7.5 and SNAP2 score > 70), which were surprisingly underrepresented among *MODY* patients and thus under negative selection during molecular evolution. We suggested and validated the first robust evidence-based thresholds, which allow improved, highly specific predictions of disease-associated GCK variations.

Glucokinase (GCK), which is one of the four mammalian isozymes that phosphorylate glucose, serves as a glucose sensor in pancreatic beta-cells, drives glucose conversion to glycogen in the liver and is expressed in the brain and endocrine cells of the gut¹. GCK is considered a key enzyme in the glycolytic pathway, particularly because of the concentration of substrate at which this enzyme shows half-maximal activity ($S_{0.5}$) that is within the physiological range of blood glucose concentrations² and because of its kinetic cooperativity, which is unique among hexokinases. In healthy people, this allows GCK to tune its response following the uptake of glucose-containing food without completely depleting blood glucose levels³.

The inactivating GCK nonsynonymous substitutions cause maturity-onset diabetes of the young (*GCK-MODY*) or insulin-deficient hyperglycemia when only one allele is affected, and they cause severe permanent neonatal diabetes mellitus (PNDM) when both alleles are inactivated, whereas the activating nonsynonymous substitutions lead to persistent hyperinsulinemic hypoglycemia of infancy (PHHI). Hundreds of nonsynonymous substitutions of the *GCK* gene have been described in many populations. The heterozygously manifested inactivating nonsynonymous substitutions are usually only associated with mild fasting hyperglycemia. Thus, many patients are not diagnosed because there is an absence of symptoms; a higher perceived prevalence of *GCK-MODY* is known in countries that perform routine blood glucose screens on pregnant women or oral glucose tolerance test (OGTT) on asymptomatic young relatives within families with multiple cases of type 2 diabetes mellitus⁴. Characterization of GCK nonsynonymous substitutions is a laborious and time-consuming task. Thus, it prevents large-scale analyses for clinical purposes, particularly when considering a structural

Charles University, Third Faculty of Medicine, Prague, Czech Republic. Correspondence and requests for materials should be addressed to P.H. (email: petr.heneberg@lf3.cuni.cz)

Variation	$S_{0.5}$ (at 5 mM ATP) [mM glucose]	n_{Hill}	$S_{0.5}$ (at 500 μ M ATP) [mM glucose]	ATP K_M (at $S_{0.5}$) [mM ATP]	ATP K_M (at 50 mM glucose) [mM ATP]	Stability [%]	k_{cat} [s^{-1}]	GlcNAc IC ₅₀ [μ M]	RAI	GSIR-T [mM glucose]
Wild type	8.82 \pm 0.07	1.72 \pm 0.04	5.65 \pm 0.22	0.36 \pm 0.00	0.47 \pm 0.01	86 \pm 1	43.8 \pm 3.4	223 \pm 19	1.00	5.0
V33A	12.77 \pm 0.26	1.56 \pm 0.05	5.75 \pm 0.21	0.53 \pm 0.03	0.63 \pm 0.02	76 \pm 3	42.6 \pm 7.4	303 \pm 14	0.51	5.9
R63S	4.44 \pm 0.17	1.64 \pm 0.04	2.75 \pm 0.08	0.23 \pm 0.02	0.35 \pm 0.02	83 \pm 0	57.7 \pm 8.8	487 \pm 35	4.17	2.9
G81D			No activity at \leq 150 mM glucose						<0.01	\geq 7.1
F150L	20.05 \pm 0.80	1.27 \pm 0.04	17.83 \pm 0.28	1.63 \pm 0.02	2.15 \pm 0.08	92 \pm 4	13.2 \pm 1.0	1470 \pm 75	0.13	6.9
T209K	9.32 \pm 0.25	1.53 \pm 0.04	7.28 \pm 0.26	0.25 \pm 0.01	0.31 \pm 0.01	88 \pm 3	25.9 \pm 5.7	293 \pm 9	0.59	5.7
R250C	7.94 \pm 0.25	1.57 \pm 0.07	4.84 \pm 0.27	0.37 \pm 0.02	0.45 \pm 0.01	81 \pm 4	36.1 \pm 6.0	287 \pm 7	0.94	5.0
M251C	55.33 \pm 1.38	1.38 \pm 0.09	Activity close to detection limits				3.30 \pm 0.15	N/D	<0.01	\geq 7.1
M251I	113.43 \pm 20.7	1.33 \pm 0.09	Activity close to detection limits				10.18 \pm 0.95	N/D	<0.01	\geq 7.1
M251V	46.4 \pm 2.11	1.59 \pm 0.03	42.36 \pm 2.24	0.46 \pm 0.02	0.51 \pm 0.00	90 \pm 5	7.6 \pm 1.7	N/D	0.01	7.1
C252R	8.37 \pm 0.19	1.61 \pm 0.02	6.52 \pm 0.37	0.27 \pm 0.01	0.34 \pm 0.02	78 \pm 3	11.8 \pm 1.9	290 \pm 16	0.27	6.2
F260L	9.07 \pm 0.19	1.54 \pm 0.02	4.76 \pm 0.05	0.36 \pm 0.01	0.41 \pm 0.01	82 \pm 3	41.8 \pm 6.5	263 \pm 24	0.95	5.1
G295D			No activity at \leq 150 mM glucose						<0.01	\geq 7.1
L314P	13.63 \pm 0.84	1.4 \pm 0.10	N/A	N/A	0.30 \pm 0.00	N/D	5.9 \pm 1.0	N/D	0.11	6.9
F316V	11.20 \pm 0.50	1.66 \pm 0.09	5.92 \pm 0.26	0.47 \pm 0.02	0.69 \pm 0.04	84 \pm 2	45.5 \pm 4.1	250 \pm 17	0.51	5.7
G318R	8.53 \pm 0.21	1.63 \pm 0.02	4.67 \pm 0.15	0.37 \pm 0.02	0.53 \pm 0.02	84 \pm 2	34.0 \pm 7.6	293 \pm 22	0.67	5.3
G385W			Activity close to detection limits				0.14 \pm 0.06	N/D	<0.01	\geq 7.1
F419L	15.00 \pm 0.32	1.67 \pm 0.01	9.57 \pm 0.21	0.33 \pm 0.02	0.43 \pm 0.02	84 \pm 3	25.4 \pm 1.1	277 \pm 19	0.19	6.6
C434Y	7.36 \pm 0.05	1.66 \pm 0.03	4.50 \pm 0.09	0.37 \pm 0.02	0.45 \pm 0.01	95 \pm 3	29.0 \pm 3.8	217 \pm 14	0.71	5.1
A454E	24.49 \pm 0.59	1.39 \pm 0.04	14.06 \pm 0.50	1.49 \pm 0.07	1.36 \pm 0.06	90 \pm 3	6.7 \pm 0.7	513 \pm 20	0.04	7.0

Table 1. The kinetic data for WT-GCK, 16 naturally occurring GCK nonsynonymous substitutions associated with MODY patients and the experimental nonsynonymous substitutions R63S, M251C and F260L. Data are shown as the means \pm SE and are representative of three to five preparations of each nonsynonymous substitution with three technical replicates analyzed for each preparation.

perspective. Recently, computational evolution- and structure-based prediction analyses were suggested to estimate the effects of particular GCK nonsynonymous substitutions⁵. These analyses aimed to identify nonsynonymous substitutions, which are likely or unlikely to have a serious impact on the protein function and stability. However, these analyses have not been paired with robust experimental data. Currently, experimental data are available based on *in vitro* kinetic analyses of over a hundred GCK nonsynonymous substitutions e.g., refs 6–9, and some nonsynonymous substitutions have been newly re-classified as non-pathogenic¹⁰.

In this study, we aimed to provide the first robust evidence for choosing the best-fit method and the evidence-based threshold to predict the effects of GCK nonsynonymous substitutions. For the first time, we compared the outcomes of prediction methods with the outcomes of *in vitro* measurements reported previously or reported newly in the course of this study, and with clinical information known from patients carrying GCK nonsynonymous substitutions. We calculated the evidence-based thresholds in order to solve the problems with negligible specificity of their previously suggested arbitrary values. By the analysis of total hypothetical GCK nonsynonymous substitutions, we predicted the effects of GCK nonsynonymous substitutions for which the clinical or *in vitro* data are still absent.

Results

***In vitro* enzyme kinetics.** We analyzed the enzyme kinetics of 16 naturally occurring GCK nonsynonymous substitutions known from MODY patients of Czech origin and the experimental nonsynonymous substitutions R63S, M251C and F260L. Five mutants – that included four naturally occurring GCK-MODY-associated mutants (G81D, M251I, G295D and G385W) and one experimental mutant (M251C) – displayed no or negligible activity at \leq 150 mM glucose. Regarding the other mutants, the GCK-MODY-associated mutants M251V, L314P, F316V, G318R and F419L demonstrated a reduced affinity for glucose that was expressed as an elevated $S_{0.5}$ measured at 5 mM ATP (one-way ANOVA $p < 0.001$, $F = 808.0$, Tukey's post-tests $p < 0.05$, excluding the enclosed means). However, when measured at the suboptimal ATP concentration (500 μ M), only a partially overlapping set of GCK-MODY-associated mutants (F150L, M251V and A454E) exhibited an elevated $S_{0.5}$ (one-way ANOVA $p < 0.001$, $F = 172.3$, Tukey's post-tests $p < 0.05$, excluding the enclosed means). There was no overlap between those differing significantly from WT-GCK at high and low levels of ATP (Table 1). The Hill coefficient differed among the tested mutants (one-way ANOVA $p < 0.001$, $F = 4.506$), but we did not find any significant differences between the mean Hill coefficient of WT-GCK and any of the mutants (all Tukey's post-tests of WT vs. mutants fell into the category of enclosed means). The ATP K_M was high particularly for F150L (i.e., four-times the control value), and it was also significantly increased in A454 and V33A (one-way ANOVA $p < 0.001$, $F = 212.3$, Tukey's post-tests $p < 0.05$, excluding the enclosed means). The tested mutants displayed a wide range of k_{cat} . Two of the MODY-associated mutants (V33A and F316V) exhibited k_{cat} values similar to WT-GCK, and all the others demonstrated decreased k_{cat} values compared to WT-GCK. Among the nonsynonymous substitutions tested, we did not find any that had a decreased stability at 30 °C (Table 1). Two nonsynonymous substitutions demonstrated

a decreased susceptibility to the competitive inhibitor of the GCK activity, *N*-acetylglucosamine (GlcNAc). The majority of the nonsynonymous substitutions did not demonstrate any change in IC_{50} of GlcNAc compared to WT-GCK, except for F150L (>6-fold higher IC_{50}) and A454E (2-fold higher IC_{50}) (Table 1).

Based on the measured data, we calculated the relative activity index (RAI) and the threshold for glucose-stimulated insulin release (GSIR-T) of each tested protein variation. While many demonstrated RAIs at or below 10% of the activity of WT-GCK, there were also MODY-associated protein variations that did not display any major difference in the RAI compared to WT-GCK. These included R250C (RAI 0.94), C434Y (RAI 0.71) and G318R (RAI 0.67). Near-normal levels of RAI resulted in mild changes of GSIR-T associated with these three nonsynonymous substitutions; the GSIR-T only ranged from 5.0 to 5.3 mM glucose. Changes in the enzyme kinetics of the other MODY-associated mutants led to changes in the RAI within the range from 5.7 to 7.1 mM glucose, which is characteristic for the GCK-MODY phenotype.

Within the variations tested, there were also three experimentally designed nonsynonymous substitutions. These included the newly identified activating mutant, R63S, which demonstrated the RAI of 4.17 and the GSIR-T of 2.9 mM and led to a marked decrease in both, $S_{0.5}$ and ATP K_M and an increased IC_{50} of GlcNAc (Table 1). Another experimental mutant was F260L, which demonstrated a neutral phenotype, reaching the RAI of 0.95, the GSIR-T of 5.1 mM, a marginally decreased $S_{0.5}$ and ATP K_M and a similar k_{cat} and IC_{50} of GlcNAc compared to WT-GCK. The last experimental mutant, M251C, exhibited a strong deactivating effect with barely detectable activity even at high doses of glucose, which led to an estimated RAI < 0.01 and GSIR-T ≥ 7.1 mM (Table 1).

Can prediction methods predict enzyme kinetics of GCK and are they in agreement with clinical data?

The prediction methods demonstrated a generally high sensitivity for deleterious MODY-associated nonsynonymous substitutions, and the sensitivity of all but one method reached at least 75%. However, many methods exhibited low sensitivity when attempting to predict the hypoglycemic phenotype, for which only PoPMuSiC 2.1 and PolyPhen-2 had sensitivity over 90%. However, high sensitivity of the latter two methods was associated with detrimental outcomes when predicting the neutral non-diabetic nonsynonymous substitutions because they reached a sensitivity of only 27 and 38%, respectively. The other prediction methods were more sensitive when predicting normoglycemic nonsynonymous substitutions and were associated with up to 94% sensitivity (SNPs&GO with GO terms excluded). However, when there was a high sensitivity for the nonsynonymous substitutions with neutral phenotypes, the false-positive ratio for detection of such nonsynonymous substitutions among those associated with disease phenotypes was increased. SNPs&GO with GO terms excluded had 28% false positive ratio for predicting nonsynonymous substitutions with the neutral phenotype. Given that the number of known normoglycemic nonsynonymous substitutions is lower by one order of magnitude compared with the disease-associated nonsynonymous substitutions of GCK, the actual number of false-positive hits suggested by SNPs&GO exceeded the number of true positive hits regardless of the inclusion of GO terms (Table 2). Previously, it was suggested to use the prediction methods in combination⁵. However, the combination of the state-of-the-art prediction methods did not lead to any improvement in the prediction of the physiologic effects of nonsynonymous substitutions when referring to the healthy vs. disease-associated status of their heterozygous carriers, particularly when one or more predictors were in disagreement (Fig. 1a).

The prediction methods demonstrated similar issues when tested against the GSIR-T values, which were calculated based on the experimentally measured enzyme kinetics data. Multiple methods had high sensitivities for both the activating and deactivating nonsynonymous substitutions, which were quantified as those with a GSIR-T lower than 4 mM or exceeding 5.5 mM, respectively. However, they failed to identify those within the normal range (4.0–5.5 mM glucose), which does not lead to PHHI, and is also below the threshold for diabetes. The most sensitive method for the neutral effects was SNPs&GO with GO terms excluded (55% sensitivity). However, the same method was associated with 29% false positive ratio for neutral predictions among nonsynonymous substitutions that cause disease-associated GSIR-T, and it was also associated with only 71% sensitivity for the deactivating nonsynonymous substitutions (Table 1). The combination of predictors was still associated with a poor ability to discriminate between a low and normal GSIR-T and with relatively low number of false-negative predictions among the nonsynonymous substitutions with a high GSIR-T (Fig. 1b).

One of the key features of GCK is its cooperativity. The majority of the tested prediction methods were able to correctly predict extreme changes in the Hill coefficient, which lead to a nearly complete loss of the cooperativity ($n_H < 1.2$) (Table 1). The combination of predictors was able to identify correctly those inducing complete or nearly complete loss of the cooperativity. However, the predictions were associated with a high number of false-positive results, and the only useful predictions of the absence of severe effects on the Hill coefficient were when seven or less algorithms agreed on the effect of the respective nonsynonymous substitution (Fig. 1c).

Reflecting the above unsatisfactory outcomes of the prediction methods, we analyzed whether the clinical phenotype of the patients can be predicted at all. First, we analyzed the relationship between the disease and experimentally measured enzyme kinetics of GCK by employing detrended correspondence analysis (DCA; Fig. 1d). The analysis involved six basic parameters for the enzyme kinetics, namely $S_{0.5}$, n_H , ATP K_M , k_{cat} , the RAI and the GSIR-T. The eigenvalues were 0.557 (axis 1), 0.111 (axis 2) and 0.026 (axis 3). At the molecular level, functional GCK variations are associated with two groups of phenotypes, the activating phenotypes (that manifest as PHHI) and the inactivating phenotypes (that manifest as GCK-MODY or PNDM). DCA distinguished between the MODY-associated and PHHI-associated nonsynonymous substitutions, and WT-GCK was positioned close to the left boundary of the MODY-associated area. The k_{cat} and RAI were responsible for most of the variability; fine resolution was allowed by the inclusion of $S_{0.5}$ and ATP K_M , which were associated with axis 2 (Fig. 1d). In contrast to the above comparison was the DCA of the nonsynonymous substitutions known from humans with the prediction methods that were suggested previously to predict the effects of GCK nonsynonymous substitutions^{5,11} (Fig. 1e). The eigenvalues were 0.053 (axis 1), 0.014 (axis 2) and 0.013 (axis 3). The prediction methods with arbitrarily set thresholds did not correctly identify the effects of nonsynonymous

Measure	Prediction method: Variable	SIFT	PolyPhen2	PhD-SNP	PoPMuSiC 2.1	SNAP2	SNPs&GO (GO terms excluded)	SNPs&GO (GO terms included)	I-Mutant 3	Align GVGD	EVmutation
Disease											
PHHI (n = 16)	Sensitivity [%]	63	94	44	94	25	13	56	81	81	75
Non-diabetic (n = 16)	Sensitivity [%]	75	38	75	27	88	94	50	20	50	N/A (53)
Non-diabetic (n = 515)	False positive ratio [%]	13.2	5.2	15.7	5.5	23.7	28.0	11.1	15.5	9.5	3.6
MODY (n = 499)	Sensitivity [%]	87.6	94.8	85.6	94.5	78.0	66.9	87.0	84.6	75.6	97.1
GSIR-T											
<4 (n = 19)	Sensitivity [%]	63	95	53	100	32	13	39	84	84	72
4–5.5 (n = 20; normal range)	Sensitivity [%]	25	10	35	15	40	55	30	20	15	5
4–5.5 (n = 106)	False positive ratio [%]	28	3	15	4	27	29	16	14	7	8
>5.5 (n = 87)	Sensitivity [%]	92	98	93	95	82	71	89	86	95	97
Hill coefficient n_H											
<1.2 (n = 23)	Sensitivity [%]	96	91	96	100	87	74	96	87	96	100
1.2–1.5 (n = 48)	Sensitivity [%]	83	96	79	94	69	50	77	88	94	94
>1.5 (n = 56; normal range)	Sensitivity [%]	18	4	23	5	36	39	23	14	11	11
>1.5 (n = 71)	False positive ratio [%]	13	4	15	4	25	30	15	13	6	4

Table 2. The summary of prediction scores for nonsynonymous substitutions in the GCK amino acid sequence, for which data were available either on their clinical phenotype, the GSIR-T or the Hill coefficient. For a detailed list of nonsynonymous substitutions analyzed, relevant raw data and references, cf. Table S2. PoPMuSiC 2.1 and I-Mutant 3 were only calculated for amino acids available in the crystal structure (PDB ID: 1V4S). Calculations for some amino acids were not available for the EVmutation. The threshold value for EVmutation was set to a median value of normoglycemic variations (-2.39) as all but two neutral nonsynonymous substitutions exceeded the originally suggested zero threshold¹¹.

substitutions as suggested by low eigenvalues and the overlap of convex hulls assigned to benign, activating and deactivating nonsynonymous substitutions (Fig. 1e). The number of predictors that correctly predicted the effect of each disease-associated nonsynonymous substitution, did not display any strong association with the clinical parameters measured in the affected patients, namely the levels of plasma glucose, glucose after 120 min OGTT, HbA_{1c}, C-peptide and age at diagnosis (Fig. S1).

In order to improve the predictions, we tested, whether there is a space for the evidence-based adjustments of the arbitrary thresholds of the prediction methods. We found that with evidence-based thresholds, three methods were capable of distinguishing between a group of disease-associated nonsynonymous substitutions and those with uncertain phenotypes. Namely, the EVmutation scores of normoglycemic nonsynonymous substitutions reached -2.39 (95% CI -3.30 – -1.32 ; Fig. 1f). When setting the threshold values to a median minus 2-times SD (EVmutation = -6.31), 43% of MODY-associated nonsynonymous substitutions (but no PHHI-associated nonsynonymous substitutions) passed this threshold applied to the outcomes of the EVmutation. Similarly, PoPMuSiC 2.1 $\Delta\Delta G$ for neutral GCK nonsynonymous substitutions reached 0.58 kcal/mol (95% CI 0.35–0.78 kcal/mol; Fig. 1g). When setting the threshold values to a median plus 2-times SD ($\Delta\Delta G = 1.42$), 42% of MODY-associated nonsynonymous substitutions (but only two PHHI-associated nonsynonymous substitution) passed this threshold applied to the outcomes of PoPMuSiC 2.1. A similarly calculated threshold for SNAP2 was 6.5 (mean -58 , 95% CI -77.06 – -39.06), which allowed to classify 75% of the tested MODY-associated (and four PHHI-associated) nonsynonymous substitutions as disease-associated (Fig. 1h). In contrast, SIFT, PolyPhen-2, I-Mutant 3 and AlignGVGD [including Grantham variation (GV) or Grantham deviation (GD) alone] did not allow any improvement in the resolution based on increases in threshold values. Other tested methods, including PhD-SNP and SNPs&GO did not allow this adjustment because they generate only binary outcomes.

We employed the evidence-based thresholds in the estimation of the theoretical frequency of putative MODY-associated variations among total hypothetical GCK nonsynonymous substitutions (Table 3). The distribution of resulting scores of EVmutation (Fig. 1i) and PoPMuSiC 2.1 (Fig. 1j) overlapped for total putative nonsynonymous substitutions and MODY-associated nonsynonymous substitutions, with slightly less MODY-associated nonsynonymous substitutions being associated with low EVmutation scores (Fig. 1i). In contrast, the total and MODY-associated SNAP2 scores did not have the same distribution, and more total nonsynonymous substitutions were associated with high SNAP2 scores (Fig. 1k). The three scores were only incompletely correlated. We found the strongest correlation between the SNAP2 and EVmutation scores (Pearson -0.778 , $p < 0.001$; Spearman 0.789, $p < 0.001$, $n = 8,189$ nonsynonymous substitutions), followed by PoPMuSiC 2.1 and

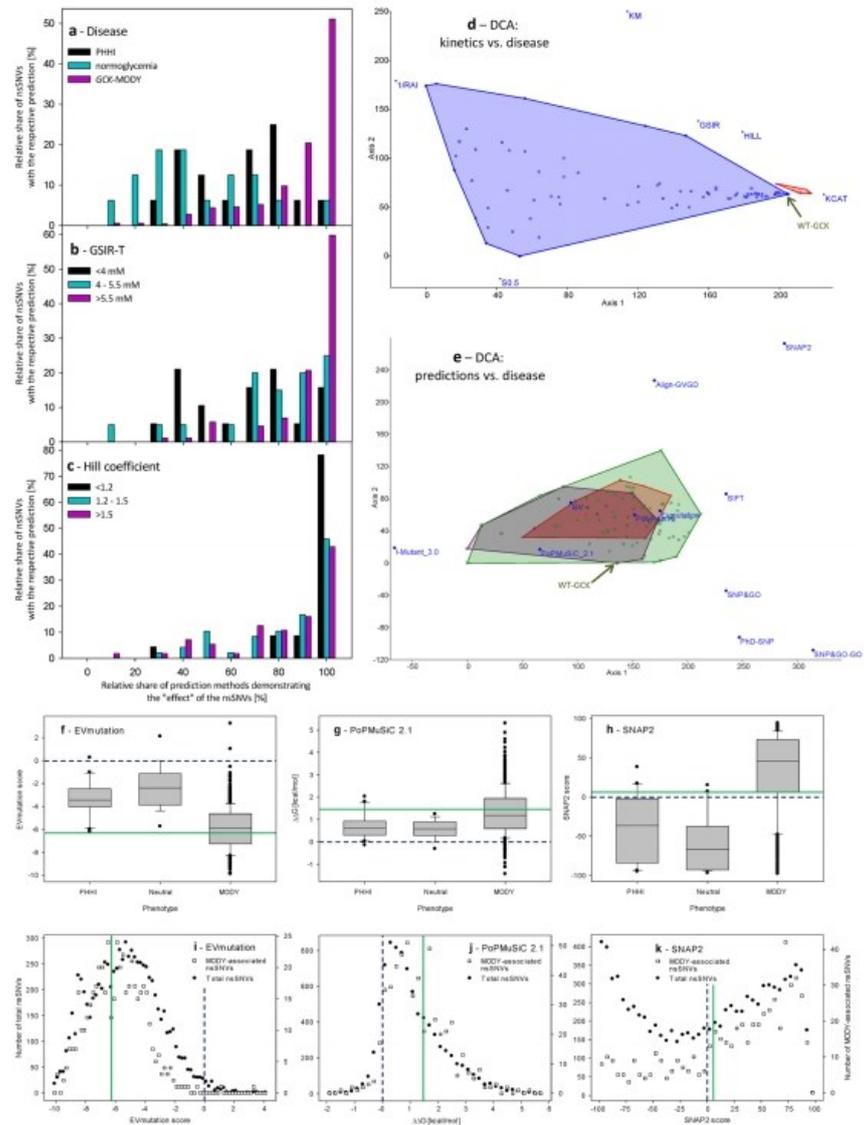


Figure 1. The efficiency of the prediction methods in predicting the effects of nonsynonymous substitutions in GCK on their enzyme kinetics and clinical phenotypes. (a–c) Number of prediction methods demonstrating the “effect” of the respective nonsynonymous substitution plotted against (a) the clinical phenotype of their heterozygous carriers, (b) the GSIR-T and (c) the Hill coefficient. (d,e) DCA comparing the nonsynonymous substitutions that were sorted according to their phenotype when plotted against (d) the outcomes of *in vitro* enzyme kinetics characterization (GCK-MODY = blue; PHHI = red polygon) and (e) predictions obtained using the state-of-the-art prediction methods (GCK-MODY = green; PHHI = red; normoglycemic = violet). Positions of WT-GCK are labeled with the green arrow. (f–k) The distribution of numerical scores of EVmutation, PoPMuSiC 2.1 and SNAP2 applied to GCK nonsynonymous substitutions with known clinical phenotype and of all putative GCK nonsynonymous substitutions. (f–h) The distribution of numerical scores of prediction methods applied to GCK nonsynonymous substitutions with known clinical phenotype. The data are shown for (f) EVmutation, (g) PoPMuSiC 2.1 and (h) SNAP2. The Tukey box plots are shown, the mean values

are presented as lines, and the 5th and 95th percentiles are displayed as symbols. (i,k) The distribution of scores of EVmutation (i), PoPMuSiC 2.1 (j) and SNAP2 (k) as calculated for total putative GCK nonsynonymous substitutions and MODY-associated nonsynonymous substitutions. The pre-set (original) thresholds are highlighted with dashed dark-blue lines and newly defined evidence-based thresholds are highlighted with solid green lines (f–k).

Method: Variable	SNAP2	PoPMuSiC 2.1	EVmutation
Number of GCK variations analyzed	8,837	8,856	8,191
Mean \pm SE	4.54 \pm 0.63	1.10 \pm 0.01	-5.35 \pm 0.03
Min	-99	-1.88	-10.15
Max	96	5.77	4.10
Median	13	0.85	-5.36
25 th percentile	-52	0.32	-7.06
75 th percentile	58	1.70	-3.79

Table 3. The estimations of the effects of total hypothetical GCK nonsynonymous substitutions. Three prediction methods, SNAP2, PoPMuSiC 2.1 and EVmutation, allowed differentiating at least in part between the neutral and MODY-associated nonsynonymous substitutions when considering their numerical outcomes. Thus, for these three methods, we computed (SNAP2 and PoPMuSiC 2.1) or retrieved (EVmutation) predictions for all possible amino acid exchanges within the GCK molecule, irrespectively on whether they are already known from humans or not.

EVmutation (Pearson -0.406 , $p < 0.001$; Spearman 0.383 , $p < 0.001$, $n = 8,493$ nonsynonymous substitutions) and SNAP2 and PoPMuSiC 2.1 (Pearson -0.362 , $p < 0.001$; Spearman 0.339 , $p < 0.001$, $n = 8,189$ nonsynonymous substitutions). When we combined the three prediction methods, the ternary transformed data allowed to distinguish the nonsynonymous substitutions associated with PPHI or normoglycemia (Fig. 2a) from those associated with MODY (Fig. 2b). The nonsynonymous substitutions with unknown clinical phenotype (so far not observed in humans) followed the same distribution pattern; most of them accumulated in the region of high SNAP2 score and low EVmutation score (Fig. 2c). Raw EVmutation and SNAP2 scores were able to differentiate nonsynonymous substitutions associated with PPHI or normoglycemia (Fig. 2d) from those associated with MODY (Fig. 2f). The distribution pattern of nonsynonymous substitutions, which were so far not observed in humans, suggests the existence of two dominant phenotypes, one considered benign (EVmutation score from -2 to -4 and SNAP2 score < -50) and the other one predicted to be highly pathogenic and surprisingly underrepresented even among MODY patients (EVmutation score < -7.5 and SNAP2 score > 70) (Fig. 2f).

Comparison between the evolutionary conservation and frequency of nonsynonymous substitutions. We found a total of 301 invariant residues (65%) out of the total 465 residues of GCK when comparing GCK protein sequences from 12 vertebrate species (Fig. S2). The large and fully evolutionarily conserved regions particularly included the residue positions: 52–66, 77–93, 140–158, 160–180, 182–217, 225–238, 248–261, 404–417 and 441–451. All known glucose binding (T168, K169, N204, D205, E256 and E290), ATP binding (T228, T332, S336, V412 and L415) and allosteric (R63, Y215, M210, Y214, V452 and V455) sites of GCK^{12,13} were fully evolutionarily conserved except for T332 (GV = 57.75) and V452 (GV = 29.61), which also displayed a high conservation (defined as $GV < 61.3$) (Table S1).

When plotted against the frequency of 1596 disease-associated families with nonsynonymous substitutions in GCK, all but one of the sites mutated in 15 or more families with MODY or PPHI ($n = 23$ amino acids) were considered as highly conserved, with zero GV score (Fig. 2g,h). The only exception was S383, which was repeatedly reported to be mutated to leucine or, less frequently, threonine in multiple European and Canadian families⁴ (Table S1). There is an overall correlation between positions that are mutated at a high frequency and sites that have a low GV (Pearson -0.182 , $p < 0.001$; Spearman -0.394 , $p < 0.001$; Fig. 2g,h). This correlation thus indicates that the nonsynonymous substitutions at highly conserved positions are strongly contributing to the disease manifestation, whereas the others may escape attention because they may not be associated with any phenotypes. Not all conserved residues were frequently mutated; for example, there were no nonsynonymous substitutions associated with the evolutionarily conserved residues 83–90. However, the extremely low frequency of disease-associated nonsynonymous substitutions was in exon 1 (any of its three forms), which corresponds to high GV values associated with the N-terminus of GCK (Fig. 2g,h).

Discussion

The use of prediction algorithms seems to be necessary part of science and diagnostics, especially in time of next-generation sequencing and other omics studies, for which the complete functional analyses of protein variations found are unfeasible and ineffective. Widely used databases list the outcomes of some of these algorithms; for example, the Ensembl genome browser provides the predictions generated by the SIFT and PolyPhen algorithms for each of the nonsynonymous substitutions listed¹⁴. This is convenient because the use of these algorithms minimizes the amount of nonsynonymous substitutions studied only to those predicted to be deleterious

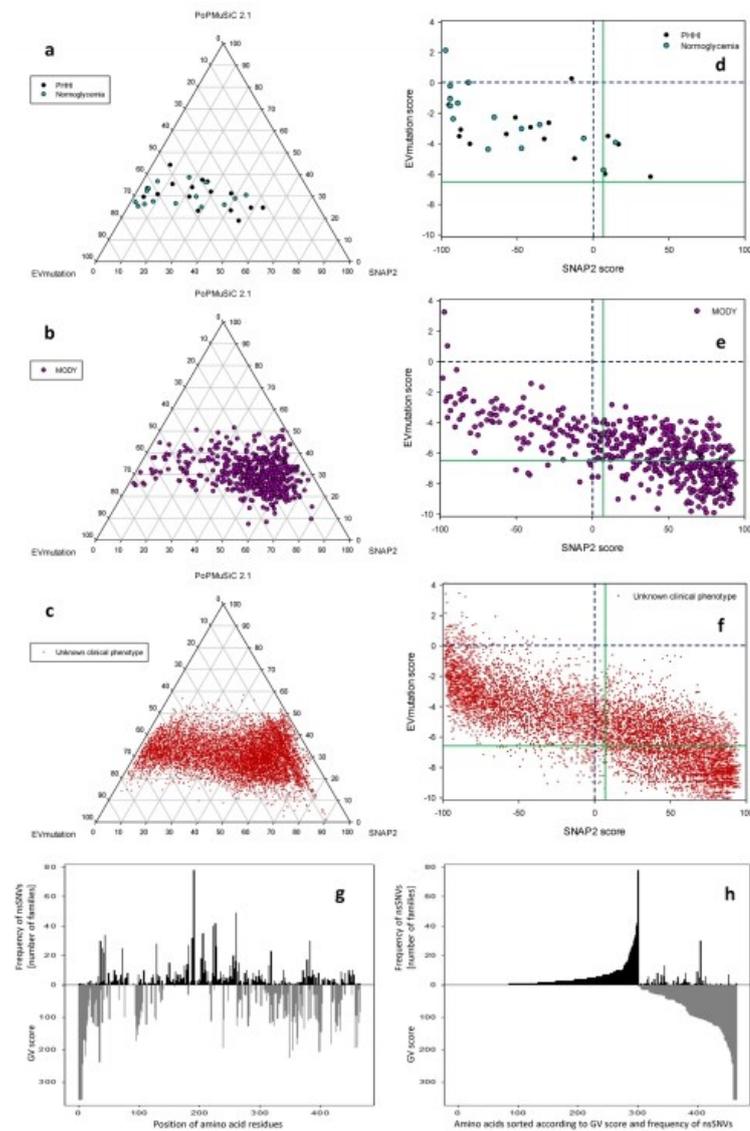


Figure 2. The combined analysis of the numerical scores of EVmutation, PoPMuSiC 2.1 and SNAP2 applied to GCK nonsynonymous substitutions with known clinical phenotype and of all putative GCK nonsynonymous substitutions and the frequency of disease-associated families. (a–c) Ternary plots of the EVmutation, PoPMuSiC 2.1 and SNAP2 scores. The numerical outcomes of each of the three prediction methods were transformed to the equal relative scale, and then ternary transformed in order to show the contributions of the predictors relative to each other, irrespectively on the absolute values of the predictions. (d–f) Scatter plots of numerical EVmutation and SNAP2 scores. The data are shown for GCK nonsynonymous substitutions associated with (a,d) PHH1 and normoglycemia, (b,e) GCK-MODY, and (c,f) with unknown clinical phenotype (so far not observed in humans). (g,h) The comparison between the GV scores and the frequency of disease-associated families with nonsynonymous substitutions in GCK. The numbers of families affected by particular nonsynonymous substitutions are based on Osbak *et al.*⁴ (data known until 2009) and from a systematic review of the literature published in 2009–2017 and listed in the Web of Science database and/or mentioned in the

HGMD database [$n = 465$ residues, of that 279 residues were disease-associated (1596 disease-associated families) and 164 residues were not evolutionarily conserved; for raw data, cf. Table S1]. Generally, areas with low GV values ($GV < 61.3$), which suggest high conservation, correspond to areas with frequently mutated residues except for S383. The data were sorted according to (g) the position of amino acid residues or (h) GV score. The pre-set (original) thresholds are highlighted with dashed dark-blue lines and newly defined evidence-based thresholds are highlighted with solid green lines (d–f).

and thus potentially causing the phenotype in the respective study subject^{15,16}. The two predictors implemented in the Ensembl genome browser are also widely used in studies focusing on particular proteins, including those that focus on the activity of GCK. Some of these studies generated data, which are in agreement with the two predictors¹⁰, but accumulating evidence suggests that they may exhibit surprisingly high false positive rates (e.g., 29% and 43%, respectively, as reported by Romeo *et al.*¹⁷) and surprisingly low rates of correct predictions (53 and 63%, respectively¹⁷). Therefore, their outcomes may be over-interpreted when used without matching the data with the measurements of enzyme kinetics and clinical data. The field of prediction methods is developing quickly, with the EVmutation method being the latest important contribution to the field¹¹. The EVmutation method reflects the epistasis by explicitly modeling interactions between all the pairs of residues in proteins, and was claimed to outperform dramatically the nowadays broadly used SIFT and PolyPhen methods¹¹. However, here we have shown that EVmutation is associated with similar issues of poor sensitivity for activating and neutral nonsynonymous substitutions as are the previously developed models, despite the sensitivity of EVmutation for deactivating nonsynonymous substitutions was similar as in other prediction methods with the best performance (Table 2). The use of evidence-based thresholds is necessary in order to avoid low selectivity (Figs 1–2). The evidence-based adjustment of the thresholds allowed confident identification of up to three quarters of MODY-associated nonsynonymous substitutions, but all the methods failed to identify selectively the nonsynonymous substitutions associated with normoglycemia or hypoglycemia. The latter two groups largely overlapped in the outcomes of all the tested prediction methods and they were also interspersed with a minority of MODY-associated variations (which, however, dominate the datasets, and thus blur any analyses of nonsynonymous substitutions associated with normoglycemia or hypoglycemia; Table 2, Figs 1–3).

When focusing on MODY-associated genes, the most authoritative work was published by Flanagan *et al.*¹⁸, who tested 66 gain-of-function and 67 loss-of-function nonsynonymous substitutions in GCK, ABCC8 and KCNJ11. They concluded that the sensitivity of SIFT and PolyPhen reached 69% and 68%, but the specificity was only 13% and 16%, respectively, with both predictors predicting more precisely the loss-of-function nonsynonymous substitutions. In another study, Rees *et al.*¹⁹ found that SIFT and PolyPhen failed to correctly predict three out of 15 nonsynonymous substitutions in GKR. False predictions of the benign phenotype in GKR by PolyPhen were also noticed by Johansen *et al.*²⁰. When focusing on GCK, most of the previous studies, which focused on MODY-associated nonsynonymous substitutions, noticed overall concordance between disease-associated phenotypes and predictions of deleterious effects based on SIFT and PolyPhen^{10,21–24}. This conclusion is in agreement with the present study, which found that these two methods tend to correctly identify MODY-associated nonsynonymous substitutions but also identify a large part of neutral nonsynonymous substitutions as deleterious and fail to correctly distinguish hypoglycemia-associated nonsynonymous substitutions (Table 2, Fig. 3).

In addition to the prediction analysis, in the present study, we provided basic kinetic characterization of 19 nonsynonymous substitutions in GCK. In agreement with previous studies^{25,26}, the dataset of MODY-associated nonsynonymous substitutions included some of the nonsynonymous substitutions, which paradoxically had near-normal kinetics. These included R250C with a GSIR-T of 5.0 mM and C434Y with a GSIR-T of 5.1 mM (Table 1). The C434Y affects one of the experimentally confirmed nitrosylation sites within the GCK molecule²⁷. Although the function of C434 nitrosylation is unknown (in contrast with the modification of C371), the association of this nonsynonymous substitution with four independent Czech families²⁸ clearly suggests its role in MODY onset and progression. Additionally, R250C is associated with a strong phenotype with manifestation during childhood and with confirmed family history²⁹, and it is known in MODY patients of Serbian and Czech origin^{29,30}. All prediction algorithms suggest its deleterious effect (Table S2). Thus, our data illustrated that the MODY-associated nonsynonymous substitutions employ various mechanisms of action. Of particular interest is the decreased sensitivity of F150L to the regulation via GlcNAc. This glucose derivative is known to inhibit glucokinase competitively in a cooperative manner³¹. It is a natural part of biopolymers on the surface of many pathogens and natural compounds in our food. Another hexokinase isoform was already confirmed to serve as a sensor for the detection of bacterial GlcNAc³²; thus, glucokinase can also serve as a pattern recognition receptor, and pathogen- and food-derived GlcNAc may differentially affect GCK action in health and disease. Such regulation would be impaired in GCK-MODY patients heterozygous for F150L. This speculation requires further verification in the near future.

In conclusion, this study provided the first robust evidence for choosing the best-fit method and the evidence-based threshold to predict the effects of GCK nonsynonymous substitutions for which *in vitro* data are still absent. Even with the newly proposed evidence-based thresholds, the precision of the available methods allowed predicting correctly up to 75% of true MODY-associated variations, leaving the remaining quarter of true MODY-associated nonsynonymous substitutions in the grey zone of uncertain predictions. The combined computational analysis of total hypothetical GCK nonsynonymous substitutions identified a group of putatively highly pathogenic variations (EVmutation score < -7.5 and SNAP2 score > 70), which were surprisingly under-represented among MODY patients. We speculate that a negative selection may play a role in the low frequency

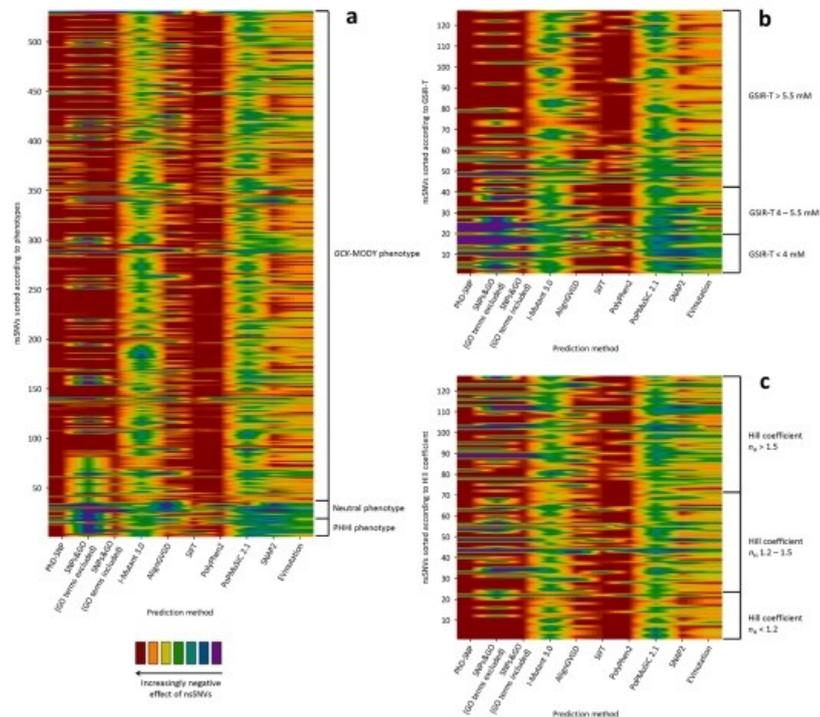


Figure 3. The prediction scores for nonsynonymous substitutions (nsSNVs) in the GCK amino acid sequence as plotted against clinical phenotypes (a), the GSIR-T (b) or the Hill coefficient (c). In the heatmaps, the nonsynonymous substitutions were sorted according to clinical phenotypes (a), the GSIR-T (b) or the Hill coefficient (c) and according to their position in the GCK molecule. The continuous scores were divided into seven categories according to computed effects of nsSNVs; the binary and ternary scores were depicted in colors of the two most extreme categories (red and violet); in case of ternary scores, the middle category was assigned the green color.

of variations predicted to be highly pathogenic despite the heterozygous manifestation of inactivating *GCK* variations is generally associated with only a relatively mild disease phenotype.

We reported here a novel approach how to interpret the outcomes of prediction methods. The tailor-made thresholds based on non-phenotypic variations can be used to improve predictions that reflect genuine effects. On a model of *GCK*, the combination of EVmutation and SNAP2 with implemented evidence-based thresholds seems to be readily applicable and feasible for the analyses of any newly found *GCK* variations. Further research should elucidate, whether the same approach can be generalized and applied across the proteome. Evidence-based cross-correlations of new bioinformatics methods with available experimental and clinical data represent the most promising approach that allows analyzing newly found amino acid sequence variations without time-consuming experiments, especially when multiple kinds of measurements would be needed in order to characterize the effect of a particular amino acid sequence variation.

Methods

We prepared a series of *GCK* constructs (Table S3) that carry previously published^{24, 28, 30, 33–35} nonsynonymous substitutions that are associated with the *GCK*-MODY phenotype in patients of Czech origin. Additionally, we prepared new constructs carrying nonsynonymous substitutions R63S, M251C and F260L for experimental purposes only. The constructs consisted of the wild-type (WT)-*GCK* isoform 1 in pGEX-5X-2 (kind gift from Dr. Navas³⁶), to which we introduced the particular nonsynonymous substitutions via site-directed mutagenesis. We measured the glucokinase activity spectrophotometrically using a coupled reaction with glucose-6-phosphate dehydrogenase, and we identified the increasing concentrations of NADPH at 340 nm as described³⁷. We computed the kinetic parameters, including the Hill coefficient n_H , the turnover number k_{cat} , $S_{0.5}$, $ATP K_M$, the protein stability, RAI and GSIR-T, as described^{37–39}. For each construct, we also performed a competitive inhibition assay using GlcNAc and calculated its IC_{50} at 5 mM glucose and 5 mM ATP.

We employed nine prediction methods (for a detailed overview, cf. Suppl. Methods.) for the prediction of phenotypic effects of GCK nonsynonymous substitutions. These included methods that use evolution-based sequence information (SIFT, PhD-SNP), as well as those that take into account the chemical and physical characteristics of amino acids (Align-GVGD) or protein structural attributes combined with multiple sequence alignment-derived information (EVMutation, PolyPhen-2, SNAP2 and SNPs&GO). A single amino acid nonsynonymous substitution can result in notable change in the protein stability, which is represented by a change in its Gibbs free energy ($\Delta\Delta G$) upon folding. Therefore, we also employed two predictors that focused on the stability properties of nonsynonymous substitutions, I-Mutant 3.0 and PoPMuSiC 2.1. We performed these predictions for an up-to-date set of published GCK-MODY- and PPHI-associated nonsynonymous substitutions and for those, which are not associated with any monogenically inherited disease effects. In addition to referring to relevant publications, we retrieved data on the GCK nonsynonymous substitutions from the Ensembl, dsSNP, UniProtKB and HGMD databases. We matched the predictions with the previously published and newly generated experimental data on enzyme kinetics and with clinical data available from previously published studies on GCK-MODY, PPHI and PNDM. Data are shown as the means \pm SE, unless stated otherwise.

Data availability. A detailed overview of the nonsynonymous substitutions analyzed, including the outcomes of the prediction methods, previously published and newly generated experimental data on enzyme kinetics, available clinical data and relevant references are all listed in Table S2; detailed protocols are provided in Suppl. Methods.

References

1. Jetton, T. L. *et al.* Analysis of upstream glucokinase promoter activity in transgenic mice and identification of glucokinase in rare neuroendocrine cells in the brain and gut. *J. Biol. Chem.* **269**, 3641–3654 (1994).
2. Lenzen, S. A fresh view of glycolysis and glucokinase regulation: history and current status. *J. Biol. Chem.* **289**, 12189–12194 (2014).
3. Larion, M. *et al.* Kinetic cooperativity in human pancreatic glucokinase originates from millisecond dynamics of the small domain. *Angew. Chem. Int. Ed.* **127**, 8247–8250 (2015).
4. Osbak, K. K. *et al.* Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. *Hum. Mutat.* **30**, 1512–1526 (2009).
5. George, D. C. *et al.* Evolution- and structure-based computational strategy reveals the impact of deleterious missense mutations on MODY 2 (maturity-onset diabetes of the young, type 2). *Theranostics* **4**, 366–385 (2014).
6. Glaser, B. *et al.* Familial hyperinsulinism caused by an activating glucokinase mutation. *N. Engl. J. Med.* **338**, 226–230 (1998).
7. Massa, O. *et al.* High prevalence of glucokinase mutations in Italian children with MODY. Influence on glucose tolerance, first-phase insulin response, insulin sensitivity and BMI. *Diabetologia* **44**, 898–905 (2001).
8. Gloyn, A. L. *et al.* Prevalence of GCK mutations in individuals screened for fasting hyperglycaemia. *Diabetologia* **52**, 172–174 (2009).
9. García-Herrero, C.-M. *et al.* Functional characterization of MODY2 mutations highlights the importance of the fine-tuning of glucokinase and its role in glucose sensing. *PLoS ONE* **7**, e30518 (2012).
10. Steele, A. M. *et al.* The previously reported T342P GCK missense variant is not a pathogenic mutation causing MODY. *Diabetologia* **54**, 2202–2205 (2011).
11. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
12. Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J. & Nagata, Y. Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure* **12**, 429–438 (2004).
13. Molnes, J. *et al.* Binding of ATP at the active site of human pancreatic glucokinase-nucleotide-induced conformational changes with possible implications for its kinetic cooperativity. *FEBS J.* **278**, 2372–2386 (2011).
14. Ensembl genome browser 88. Available from <http://www.ensembl.org/> (2017).
15. Flannick, J. *et al.* Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat. Genet.* **45**, 1380–1385 (2013).
16. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).
17. Romeo, S. *et al.* Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J. Clin. Invest.* **119**, 70–79 (2009).
18. Flanagan, S. E., Patch, A. M. & Ellard, S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomarkers* **14**, 533–537 (2010).
19. Rees, M. G. *et al.* Correlation of rare coding variants in the gene encoding human glucokinase regulatory protein with phenotypic, cellular, and kinetic outcomes. *J. Clin. Invest.* **122**, 205–217 (2012).
20. Johansen, C. T., Wang, J. & Lanktree, M. B. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42**, 684–687 (2010).
21. Beer, N. L. *et al.* Insights into the pathogenicity of rare missense GCK variants from the identification and functional characterization of compound heterozygous and double mutations inherited in cis. *Diabetes Care* **35**, 1482–1484 (2012).
22. Kanthimathi, S. *et al.* Glucokinase gene mutations (MODY 2) in Asian Indians. *Diabetes Technol. Therap.* **16**, 180–185 (2014).
23. Estalella, I. *et al.* Mutations in GCK and HNF-1 α explain the majority of cases with clinical diagnosis of MODY in Spain. *Clin. Endocrinol.* **67**, 538–546 (2007).
24. Valentinová, L. *et al.* Identification and functional characterization of novel glucokinase mutations causing maturity-onset diabetes of the young in Slovakia. *PLoS ONE* **7**, e34541 (2012).
25. Sagen, J. V. *et al.* From clinicogenetic studies of maturity-onset diabetes of the young to unraveling complex mechanisms of glucokinase regulation. *Diabetes* **55**, 1713–1722 (2006).
26. Gloyn, A. L. *et al.* Insights into the structure and regulation of glucokinase from a novel mutation (V62M), which causes maturity-onset diabetes of the young. *J. Biol. Chem.* **280**, 14105–14113 (2005).
27. Rizzo, M. A. & Piston, D. W. Regulation of β cell glucokinase by S-nitrosylation and association with nitric oxide synthase. *J. Cell Biol.* **161**, 243–248 (2003).
28. Pruhova, S. *et al.* Glucokinase diabetes in 103 families from a country-based study in the Czech Republic: geographically restricted distribution of two prevalent GCK mutations. *Pediatr. Diabetes* **11**, 529–535 (2010).
29. Milenković, T., Zdravković, D. & Mitrović, K. [Novel glucokinase mutation in a boy with maturity-onset diabetes of the young]. *Srp. Arh. Celok. Lek.* **136**, 542–544 (2008).
30. Pinterova, D. *et al.* Six novel mutations in the GCK gene in MODY patients. *Clin. Genet.* **71**, 95–96 (2007).
31. Cárdenas, M. L., Rabajlle, E. & Niemyer, H. Suppression of kinetic cooperativity of hexokinase D (glucokinase) by competitive inhibitors. A slow transition model. *Eur. J. Biochem.* **145**, 163–171 (1984).
32. Wolf, A. J. *et al.* Hexokinase is an innate immune receptor for the detection of bacterial peptidoglycan. *Cell* **166**, 624–636 (2016).

33. Lukášová, P. *et al.* Screening of mutations and polymorphisms in the glucokinase gene in Czech diabetic and healthy control populations. *Physiol. Res.* 57, S99–S108 (2008).
34. Pruhova, S. *et al.* Genetic epidemiology of MODY in the Czech republic: new mutations in the MODY genes *HNF-4 α* , *GCK* and *HNF-1 α* . *Diabetologia* 46, 291–295 (2003).
35. Urbanová, J. *et al.* Positivity for islet cell autoantibodies in patients with monogenic diabetes is associated with later diabetes onset and higher HbA_{1c} level. *Diabet. Med.* 31, 466–471 (2014).
36. García-Herrero, C. M. *et al.* Functional analysis of human glucokinase gene mutations causing MODY2: exploring the regulatory mechanisms of glucokinase activity. *Diabetologia* 50, 325–333 (2007).
37. Davis, E. A. *et al.* Mutants of glucokinase cause hypoglycaemia- and hyperglycaemia syndromes and their analysis illuminates fundamental quantitative concepts of glucose homeostasis. *Diabetologia* 42, 1175–1186 (1999).
38. Matschinsky, F. M. Assessing the potential of glucokinase activators in diabetes therapy. *Nat. Rev. Drug Discov.* 8, 399–416 (2009).
39. Matschinsky, F. M. *et al.* The glucokinase system and the regulation of blood sugar. In Matschinsky, D. M. & Magnuson, M. A., Eds Molecular pathogenesis of MODYs. Basel, Karger, pp. 99–108 (2000).

Acknowledgements

We thank Maria Angeles Navas (Complutense University of Madrid) for the WT-GCK construct, Michal Anděl, Blanka Rypáčková, Jitka Tomešová and Jana Urbanová for sharing the data on the newly identified GCK-MODY patients and Michal Boušek for expert technical assistance. The study was supported by the Czech Science Foundation project 15-03834Y and Charles University in Prague projects Primus/MED/32 and 260387/SVV/2017. All financial support for the work was acknowledged.

Author Contributions

P.H. and D.S. conceived and designed the experiments, analyzed the data, wrote the paper and are responsible for the integrity of this work. D.S., L.K., K.V. and M.T. acquired data. All authors revised the article's intellectual content and approved the final version.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-09810-0

Competing Interests: The presentation of work-in-progress data was supported by Eli Lilly. The authors declare that they have no other conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

Daniela Šimčíková, Petr Heneberg

Identification of alkaline pH optimum of human glucokinase because of ATP-mediated bias correction in outcomes of enzyme assays

Scientific Reports (2019) 9: 11422

Abstract:

Adenosine triphosphate (ATP) is a crucial substrate and energy source commonly used in enzyme reactions. However, we demonstrated that the addition of this acidic compound to enzyme assay buffers can serve as a source of unnoticed pH changes. Even relatively low concentrations of ATP (up to 5 mM) shifted pH of reaction mixtures to acidic values. For example, Tris buffer lost buffering capacity at pH 7.46 by adding ATP at a concentration higher than 2 mM. In addition to the buffering capacity, the pH shifts differed with respect to the buffer concentration. High ATP concentrations are commonly used in hexokinase assays. We demonstrated how the presence of ATP affects pH of widely used enzyme assay buffers and inversely affected KM of human hexokinase 2 and $S0.5$ of human glucokinase. The pH optimum of human glucokinase was never reported before. We found that previously reported optimum of mammalian glucokinase was incorrect, affected by the ATP-induced pH shifts. The pH optimum of human glucokinase is at pH 8.5–8.7. Suggested is the full disclosure of reaction conditions, including the measurement of pH of the whole reaction mixtures instead of measuring pH prior to the addition of all the components.

OPEN

Identification of alkaline pH optimum of human glucokinase because of ATP-mediated bias correction in outcomes of enzyme assays

Daniela Šimčíková & Petr Heneberg 

Adenosine triphosphate (ATP) is a crucial substrate and energy source commonly used in enzyme reactions. However, we demonstrated that the addition of this acidic compound to enzyme assay buffers can serve as a source of unnoticed pH changes. Even relatively low concentrations of ATP (up to 5 mM) shifted pH of reaction mixtures to acidic values. For example, Tris buffer lost buffering capacity at pH 7.46 by adding ATP at a concentration higher than 2 mM. In addition to the buffering capacity, the pH shifts differed with respect to the buffer concentration. High ATP concentrations are commonly used in hexokinase assays. We demonstrated how the presence of ATP affects pH of widely used enzyme assay buffers and inversely affected K_M of human hexokinase 2 and $S_{0.5}$ of human glucokinase. The pH optimum of human glucokinase was never reported before. We found that previously reported optimum of mammalian glucokinase was incorrect, affected by the ATP-induced pH shifts. The pH optimum of human glucokinase is at pH 8.5–8.7. Suggested is the full disclosure of reaction conditions, including the measurement of pH of the whole reaction mixtures instead of measuring pH prior to the addition of all the components.

Enzyme assays are an integral part of research, since enzymes are fundamental for maintaining the life functions of organisms. In the course of the development of enzyme assays, we must overcome many obstacles, since every enzyme represents its own individuality, which manifests by different requirements of storage, expression and purification processes, pH and temperature stability. Therefore, an adequate and well-defined environment for every enzyme is needed in order to accurately interpret outcomes of individual enzyme assays.

Many enzyme assays possess a high potential of biased outcomes caused by an inaccurately defined pH of the reaction. The well-defined pH of the reaction depends on both a reaction buffer and other components that are necessary for the reaction to proceed¹. In this regard, the addition of an acidic substance, such as adenosine triphosphate (ATP), can serve as a source of unnoticed pH changes in enzymatic reactions. For example, high ATP concentrations are characteristic of hexokinase (HK) assays, where they have the potential to affect the Michaelis constant K_M of HKs as well as the $S_{0.5}$ of glucokinase (GCK). Both constants, K_M and $S_{0.5}$, define the values of the substrate concentration at which the rate of reaction is half of the limiting rate. Both enzymes have been subjected to pharmacological studies, in which interpretations of the substrate or the inhibitor/activator affinity and imitation of physiological conditions played a prominent role^{2–4}. Surprisingly, many studies on HKs and other enzymes provide little information regarding the use of ATP in their reaction mixtures, or they often disclose pH of the buffer that has been measured only before the subsequent addition of other components, including ATP (Tables S1, S2). Moreover, the pH optimum of human GCK was never determined experimentally and was only inferred from seminal studies by Salas *et al.*⁵ that used rabbit GCK instead of its human ortholog (Table S1).

HKs appear to be sensitive to pH-induced changes. Available HK structures suggest that their catalytic domains possess deep crevice between the large and small sub-domains. Glucose binds to the bottom of this crevice using hydrogen bonds to both the subdomains and their connecting region. The interface between the

Charles University, Third Faculty of Medicine, Prague, Czech Republic. Correspondence and requests for materials should be addressed to P.H. (email: petr.heneberg@lf3.cuni.cz)

Received: 1 November 2018

Accepted: 8 July 2019

Published online: 06 August 2019

two subdomains is rich in acidic residues and, therefore, susceptible to pH-mediated regulation⁶. Only relatively few studies on pH kinetics of HKs were previously published; particularly scarce are data on the impact of pH on enzyme activity due to protonation/deprotonation of active site residues. In this regard, the study of calorimetric profiles of yeast HK A revealed a single thermal transition in the acidic pH and two independent transitions in the alkaline pH⁶. One of the transitions at the alkaline pH corresponds to the small subdomain and the second transition at the alkaline pH corresponds to the large subdomain. The ionization state of the acidic residues at the active site likely regulates domain movements, induce the cleft closure and therefore cause the inaccessibility of active site to glucose⁶. In bovine brain HK (i.e., HK1), the $-\log V_i$ and $-\log V_i/K_M$ profiles displayed slopes of -1 when testing their pH kinetics with glucose and Mg ATP as substrate⁷. This is consistent with the protonation of a single group on the enzyme. Two ionizable groups were suggested to be involved in the reaction, one employed in the ATP binding and catalysis, and another playing a role in glucose binding⁷.

In the present study, we focus on enzyme assays, in which the outcomes are sensitive to the pH of the reaction. As a proof of principle, we investigated HK assays, in which the addition of ATP can serve as a source of unnoticed pH changes in enzymatic reactions. We showed how ATP concentration affects pH of different buffers and demonstrated the influence of pH on changes to the K_M values of human HK2 as well as the $S_{0.5}$ values of human GCK.

Methods

To test the buffering capacity of commonly used enzyme assay buffers according to changing ATP concentrations, we prepared the reaction mixtures as follows: 1 mL of the respective buffer; 0.4 mL of the GST-GCK elution buffer, with or without the tested enzyme; 0.1 M ATP in various volumes; and dH₂O added to adjust the total volume to 2 mL. The composition of the elution buffer was as follows: 2.6 mM NADP, 0.1 mL 1 M glucose, 0.2 mL 50 mM Tris, 200 mM KCl, and 5 mM DTT; pH adjusted to 8.0. We kept all solutions at 23 °C, except for ATP and NADP, which were kept on ice. In some cases, we observed the shift in pH towards more acidic values after the addition of NADP. The amount of NADP was constant in all mixtures; therefore, the observed changes in pH were caused only by changing ATP concentration. The ATP solution was added to the buffers in a form of a 100 mM aqueous solution that was prepared directly from the ATP disodium salt hydrate powder, without any adjustment of its pH and without the addition of any salts. ATP was always added shortly before the experiments to avoid any potential issues with its stability. All chemicals were purchased from Sigma-Aldrich (St. Louis, MO).

We prepared GST-GCK as described previously^{8,9} by a one-step purification using GSTrap HP (GE Healthcare, Chicago, IL). In the case of HK2, we introduced the insert encoding HK2 into pET-28a(+) and expressed HK2 in BL21(DE3)pLysS *E. coli*. We induced HK2 expression by the addition of 1 mM IPTG and subsequently cultivated the cells for 16 h at 22 °C. Afterwards, we purified HK2 using HisTrap HP (GE Healthcare, Chicago, IL). We measured the pH optimum of the purified GST-GCK using a coupled reaction with glucose-6-phosphate dehydrogenase as described previously¹⁰. We conducted measurements at 1 mM ATP, 50 mM glucose, 100 mM Tris, for pH range of 7.5–8.8 or 100 mM glycine for pH range of 8.6–10.3. We measured HK2 and GST-GCK activity using a coupled reaction with glucose-6-phosphate dehydrogenase as described previously^{10,11}. In the case of HK2, we measured enzymatic activity in the range of 0–2 mM glucose, unlike GST-GCK, in which the activity we measured was in the range of 0–150 mM. We prepared all the buffers and measured the enzyme kinetics at 23 °C, thereby excluding effects of temperature on pH of the solutions used.

We performed all measurements in three or more independent experiments, each performed in triplicate. We analyzed the obtained curves by nonlinear regression using SigmaPlot 12.0.

Results and Discussion

Even relatively low concentrations of ATP (up to 5 mM) shifted pH of reaction mixtures to acidic values (Fig. 1). The ATP-induced pH shifts differed with respect to the buffering capacity, which is maximal in pK_a of the respective substance, with Tris having its pK_a at 8.07 and glycine at 9.8. Tris buffer lost its buffering capacity at pH 7.46 by adding ATP at a concentration higher than 2 mM (Fig. 1A). Similarly, glycine buffer lost its buffering capacity with the increasing difference from pK_a of glycine (Fig. 1C,D). In addition to the buffering capacity, the pH shifts differed with respect to the buffer concentration, with lower buffer concentrations leading to more prominent ATP-induced pH shifts (Fig. 1B).

To demonstrate that the ATP-induced pH shifts may severely affect outcomes of enzyme kinetics measurements, we investigated HK assays, in which high ATP concentrations are commonly used. First, we re-evaluated the pH optimum of GCK that was previously published by Salas *et al.*⁵ (Fig. 1E) based on the rabbit GCK ortholog. This is still the only reference curve of the pH optimum of mammalian GCK despite the issues reported below; the pH optimum of human GCK itself was never tested. We particularly contradicted the use of glycine buffer in the published pH range, since a pH lower than 8.0 is outside of the reliable glycine buffering range. Thus, the addition of 5 mM ATP increased the pH of the reaction mixture outside of the desired pH values (Fig. 1C). Therefore, we measured the pH optimum of human GCK in buffers that were devoid of these issues (200 mM Tris and 200 mM glycine) and reflected the pH of the whole reaction mixtures (Fig. 1F), setting the pH optimum of human GCK to pH 8.5–8.7, high above physiological intracellular pH.

The pH optimum higher than physiological intracellular pH of adult differentiated cells is already known from some other glycolytic enzymes and is likely related to increased glycolysis in cancer cells. Similarly to the present data on GCK (Fig. 1), phosphofructokinase, a key rate-limiting glycolysis enzyme, has pH optimum at values higher than those considered physiological intracellular pH of adult differentiated cells. Moreover, the activity of phosphofructokinase drops by over two orders of magnitude when the pH decreases from 7.5 to 7.0^{12,13}. Phosphofructokinase studies revealed that its inhibition by protons is allosterically modified by ATP, and the inhibitory effects are lower when lower ATP concentrations are present. The inhibitory effects of ATP at low pH can be reversed by increased concentrations of AMP^{14–18}. Another glycolytic enzyme with higher than expected pH optimum is lactate dehydrogenase, which regenerates NAD⁺ for glycolysis, and which has its pH optimum at pH 7.5¹⁹. Consistent with the above findings, the acidosis is well known to decrease glycolytic activity both *in vitro* and *in vivo*^{20–22}. Importantly, tumors have been repeatedly

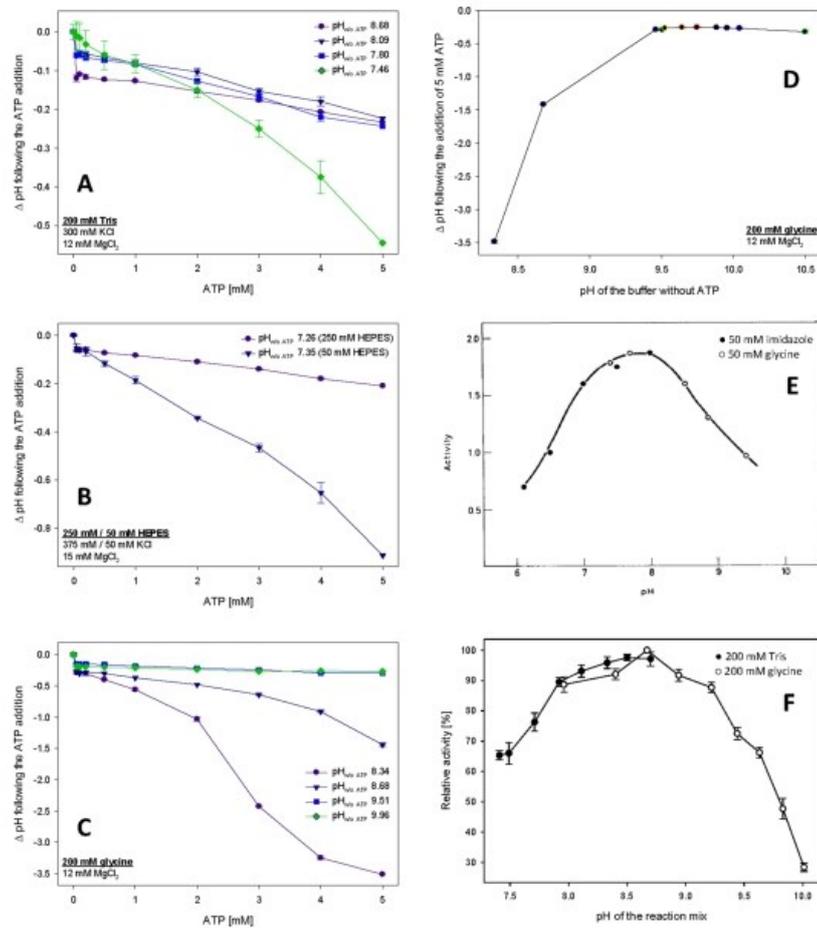


Figure 1. ATP addition affects pH of enzyme assay buffers and affects the measurement of the pH optimum as demonstrated in the example of GCK. (A) Effects of ATP addition to the buffer composed of 200 mM Tris, 300 mM KCl and 12 mM MgCl₂, pH 7.46, 7.80, 8.09 and 8.68 as measured prior to the ATP addition, and tested before and after the addition of up to 5 mM ATP. (B) Effects of buffer concentration on buffer capacity demonstrated as effects of ATP addition to the buffer containing either 250 mM HEPES, 375 mM KCl and 15 mM MgCl₂ or 50 mM HEPES, 50 mM KCl and 15 mM MgCl₂, pH 7.26 and 7.35, respectively, as measured prior to the ATP addition and tested before and after the addition of up to 5 mM ATP. (C) Effects of ATP addition to the buffer composed from 200 mM glycine and 12 mM MgCl₂, pH 8.34, 8.68, 9.51 and 9.96 as measured prior to the ATP addition and tested before and after the addition of up to 5 mM ATP. (D) Increase in buffering capacity of the glycine buffer demonstrated as the decrease in Δ pH of 200 mM glycine and 12 mM MgCl₂ following the increase of pH of the buffer without ATP closer to its pK_a. The demonstrated pH range reflects the range of the use of glycine buffer by Salas *et al.*⁵. (E) The pH optimum curve of mammalian GCK reprinted with permission from Salas *et al.*⁵. Note the use of glycine buffer in the range where it is out of its buffering capacity. Republished with permission of American Society for Biochemistry and Molecular Biology, from Salas, J., Salas, M., Vinuela, E. & Sols, A.: Glucokinase of rabbit liver, *J. Biol. Chem.* 240, edition 1, 1965, pp. 1014–1018; permission conveyed through Copyright Clearance Center, Inc. (F) The pH optimum curve of human GCK generated during the course of the present study in reaction buffers containing either 200 mM Tris (pH range up to 8.5) or 200 mM glycine (pH range from 8.1).

reported to have a higher intracellular (and lower extracellular) pH compared to normal differentiated adult cells^{23,24}. The intracellular pH of cancer cells is usually between 7.3 and 7.6, whereas the normal differentiated adult cells have an intracellular pH ~7.2²⁴, which applies to cancer cells from their early developmental stages²⁵, is sufficient to induce

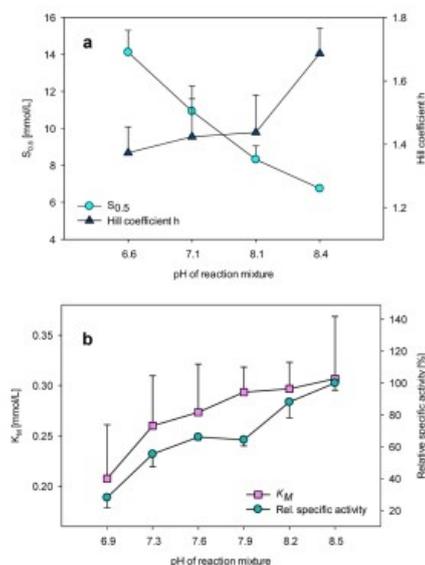


Figure 2. The effects of the pH on enzyme kinetics of human GCK and HK2. (a) $S_{0.5}$ and the Hill index (h) of human GCK as measured in the pH range from 6.6 to 8.4. (b) K_M and relative specific activity of human HK2 as measured in the pH range from 6.9 to 8.5.

dysplasia²⁶, and is further facilitated during their neoplastic progression²⁷. The role of the dynamics of intracellular pH remains in regulating cell fate decisions and cancer progression remains understudied²⁸ as the previous biochemical studies were only recently supported by an emerging evidence of transient changes in intracellular pH during key cellular processes, including cell cycle progression²⁹, migration^{30,31}, or differentiation^{32,33}. As a large part of previously reported data on HK kinetics is likely affected by the artifacts introduced by the addition of ATP, trusting in the buffering capacity of the respective buffers, we next focused on whether the effects of the pH changes could be generalized to all HKs. We compared the data obtained using human GCK with those obtained using HK2. The $S_{0.5}$ of GCK gradually decreased by over one half with a pH increase from 6.6 to 8.4 (Fig. 2a). In contrast, the K_M value of HK2 gradually increased by one half with a pH increase from 6.9 to 8.5 (Fig. 2b). During the same treatment, the cooperativity of GCK remained stable or slightly increased at the highest pH interval that was tested (Fig. 2a). In contrast, the relative specific activity of HK2 was stable only in a narrow interval of pH 7.3–7.9, while it decreased substantially at pH 6.9 and increased at pH 8.2 and higher (Fig. 2b). Collectively, these data suggest that HKs do not respond to pH changes uniformly. Their pH-dependent changes in activity may be of particular importance in emerging role of HKs in dysregulated cancer metabolism^{34,35}, including their involvement in pH-induced changes of cancer metabolism^{36–38}.

We proved that even low concentrations of reaction components, such as ATP, may shift the pH of the reaction to undesired values and adversely influence outcomes. We suggest paying more attention to the choice of appropriate assay buffer, its concentration, buffering capacity and the influence of other substances needed for enzyme assays, since enzymes are very susceptible to pH changes. We demonstrated the problem of the sufficient choice of buffer on the pH optimum of GCK, previously measured by Salas *et al.*⁵. Due to the consideration of the pH of the whole reaction, we identified a more alkaline pH optimum of human GCK (pH 8.5–8.7) compared to the previously reported optimum of rabbit GCK (pH 7.5–8.0), which was measured under likely irreproducible conditions⁵. We must admit that the effect of ATP could be eliminated based on the use of a prebuffered ATP solution, which can be purchased or prepared homemade. Nevertheless, the use of such a solution instead of pure ATP has not been mentioned in any of the publications reporting HK assays.

The topic of pH reliability, biased assay outcomes and confusing descriptions of enzyme conditions belong to underestimated but important issues related to research integrity and reproducibility. As an example of a good practice, we would like to cite F. M. Matschinsky and colleagues¹⁰, who described the GCK assay as follows: “Glucokinase activity was measured spectrophotometrically using an NADP⁺ coupled assay with glucose-6-phosphate dehydrogenase as described³⁹. The pH of all assays was 7.4, except for assays which assessed the inhibition of glucokinase by glucokinase regulatory protein (GKRP) where a pH of 7.1 was used.” Although Matschinsky and colleagues referred to a previous paper with regards to the method used, they completed the description of the enzyme assay with information about the pH. Notwithstanding, the common practice is rather to introduce only individual components of the reaction mixture (Table S1). Therefore, we suggest the practice of a full disclosure of reaction conditions of the experiments, including the measurement of the pH of the whole reaction mixtures.

Data Availability

All data are available in the main text or in the supplementary materials. Figure 1E is reprinted with permission from Salas *et al.*⁵.

References

1. Ferreira, C. M., Pinto, I. S., Soares, E. V. & Soares, H. M. Un)suitability of the use of pH buffers in biological, biochemical and environmental studies and their interaction with metal ions – a review. *RSC Adv.* **5**, 30989–31003 (2015).
2. Lin, H. *et al.* Discovery of a novel 2,6-disubstituted glucosamine series of potent and selective hexokinase 2 inhibitors. *ACS Med. Chem. Lett.* **7**, 217–222 (2015).
3. Zhang, H. N. *et al.* Systematic identification of arsenic-binding proteins reveals that hexokinase-2 is inhibited by arsenic. *Proc. Natl. Acad. Sci. USA* **112**, 15084–15089 (2015).
4. Fujieda, H. *et al.* Discovery of a potent glucokinase activator with a favorable liver and pancreas distribution pattern for the treatment of type 2 diabetes mellitus. *Eur. J. Med. Chem.* **156**, 269–294 (2018).
5. Salas, J., Salas, M., Vinuela, E. & Sols, A. Glucokinase of rabbit liver. *J. Biol. Chem.* **240**, 1014–1018 (1965).
6. Kumar, D. P., Tiwari, A. & Bhat, R. Effect of pH on the stability and structure of yeast hexokinase A. Acidic amino acid residues in the cleft region are critical for the opening and the closing of the structure. *J. Biol. Chem.* **279**, 32093–32099 (2004).
7. Solheim, L. P. & Fromm, H. J. pH kinetic studies of bovine brain hexokinase. *Biochemistry* **19**, 6074–6080 (1980).
8. Šimčíková, D., Kocková, L., Vackářová, K., Tešínský, M. & Heneberg, P. Evidence-based tailoring of bioinformatics approaches to optimize methods that predict the effects of nonsynonymous amino acid substitutions in glucokinase. *Sci. Rep.* **7**, 9499 (2017).
9. García-Herrero, C. M. *et al.* Functional analysis of human glucokinase gene mutations causing MODY2: exploring the regulatory mechanisms of glucokinase activity. *Diabetologia* **50**, 325–333 (2007).
10. Davis, E. A. *et al.* Mutants of glucokinase cause hypoglycaemia- and hyperglycaemia syndromes and their analysis illuminates fundamental quantitative concepts of glucose homeostasis. *Diabetologia* **42**, 1175–1186 (1999).
11. Nawaz, M. H. *et al.* The catalytic inactivation of the N-half of human hexokinase 2 and structural and biochemical characterization of its mitochondrial conformation. *Biosci. Rep.* **38**, BSR20171666 (2018).
12. Fidelman, M. L., Seeholzer, S. H., Walsh, K. B. & Moore, R. D. Intracellular pH mediates action of insulin on glycolysis in frog skeletal muscle. *Am. J. Physiol.* **124**, 87–93 (1982).
13. Ui, M. A role of phosphofructokinase in pH-dependent regulation of glycolysis. *Biochim. Biophys. Acta* **124**, 310–322 (1966).
14. Trivedi, B. & Danforth, W. H. Effect of pH on the kinetics of frog muscle phosphofructokinase. *J. Biol. Chem.* **241**, 4110–4112 (1966).
15. Dobson, G. P., Yamamoto, E. & Hochachka, P. W. Phosphofructokinase control in muscle: nature and reversal of pH-dependent ATP inhibition. *Am. J. Physiol.* **250**, R71–R76 (1986).
16. Ercińska, M., Deas, J. & Silver, I. A. The effect of pH on glycolysis and phosphofructokinase activity in cultured cells and synaptosomes. *J. Neurochem.* **65**, 2765–2772 (1995).
17. Frieden, C., Gilbert, H. R. & Bock, P. E. Phosphofructokinase III. Correlation of the regulatory kinetic and molecular properties of the rabbit muscle enzyme. *J. Biol. Chem.* **251**, 5644–5647 (1976).
18. Andrés, V., Carreras, J. & Cussó, R. Regulation of muscle phosphofructokinase by physiological concentrations of bisphosphorylated hexoses: effect of alkalization. *Biochem. Biophys. Res. Commun.* **172**, 328–334 (1990).
19. Gray, J. A. Kinetics of enamel dissolution during formation of incipient caries-like lesions. *Arch. Oral Biol.* **11**, 397–422 (1966).
20. Seglen, P. O. The effect of perfusate pH on respiration and glycolysis in the isolated rat liver perfused with an erythrocyte- and protein-free medium. *Biochim. Biophys. Acta* **264**, 398–410 (1972).
21. Wu, T. F. & Davis, E. J. Regulation of glycolytic flux in an energetically controlled cell-free system: the effects of adenine nucleotide ratios, inorganic phosphate, pH, and citrate. *Arch. Biochem. Biophys.* **209**, 85–89 (1981).
22. Folbergrová, J., MacMillan, V. & Siesjö, B. K. The effect of hypercapnic acidosis upon some glycolytic and Krebs cycle-associated intermediates in the rat brain. *J. Neurochem.* **19**, 2507–2517 (1972).
23. Webb, B. A., Chimenti, M., Jacobson, M. P. & Barber, D. L. Dysregulated pH: a perfect storm for cancer progression. *Nat. Rev. Cancer* **11**, 671–677 (2011).
24. White, K. A., Grillo-Hill, B. K. & Barber, D. L. Cancer cell behaviors mediated by dysregulated pH dynamics at a glance. *J. Cell Sci.* **130**, 663–669 (2017).
25. Reshkin, S. J. *et al.* Na⁺/H⁺ exchanger-dependent intracellular alkalization is an early event in malignant transformation and plays an essential role in the development of subsequent transformation-associated phenotypes. *FASEB J.* **14**, 2185–2197 (2000).
26. Grillo-Hill, B. K., Choi, C., Jimenez-Vidal, M. & Barber, D. L. Increased H⁺ efflux is sufficient to induce dysplasia and necessary for viability with oncogene expression. *eLife* **4**, e03270 (2015).
27. Cardone, R. A. *et al.* A novel NHE1-centered signaling cassette drives epidermal growth factor receptor-dependent pancreatic tumor metastasis and is a target for combination therapy. *Neoplasia* **17**, 155–166 (2015).
28. Tatapudy, S., Aloisio, F., Barber, D. & Nystul, T. Cell fate decisions: emerging roles for metabolic signals and cell morphology. *EMBO Rep.* **18**, 2105–2118 (2017).
29. Putney, L. K. & Barber, D. L. Na-H exchange-dependent increase in intracellular pH times G2/M entry and transition. *J. Biol. Chem.* **278**, 44645–44649 (2003).
30. Denker, S. P. & Barber, D. L. Cell migration requires both ion translocation and cytoskeletal anchoring by the Na-H exchanger NHE1. *J. Cell Biol.* **159**, 1087–1096 (2002).
31. Stock, C. & Schwab, A. Protons make tumor cells move like clockwork. *Pflugers Arch.* **458**, 981–992 (2009).
32. Ulmschneider, B. *et al.* Increased intracellular pH is necessary for adult epithelial and embryonic stem cell differentiation. *J. Cell Biol.* **215**, 345–355 (2016).
33. Singh, Y. *et al.* Alkaline cytosolic pH and high sodium hydrogen exchanger 1 (NHE1) activity in Th9 cells. *J. Biol. Chem.* **291**, 23662–23671 (2016).
34. Hay, N. Reprogramming glucose metabolism in cancer: can it be exploited for cancer therapy? *Nat. Rev. Canc.* **16**, 635–649 (2016).
35. Wang, L. *et al.* Hexokinase 2-mediated Warburg effect is required for PTEN- and p53-deficiency-driven prostate cancer growth. *Cell Rep.* **8**, 1461–1474 (2014).
36. Quach, C. H. *et al.* Mild alkalization acutely triggers the Warburg effect by enhancing hexokinase activity via voltage-dependent anion channel binding. *PLoS ONE* **11**, e0159529 (2016).
37. Harduindey, S. *et al.* Cellular acidification as a new approach to cancer treatment and to the understanding and therapeutics of neurodegenerative diseases. *Semin. Canc. Biol.* **43**, 157–179 (2017).
38. Hardonniere, K., Huc, L., Sergent, O., Holme, J. A. & Lagadic-Gossmann, D. Environmental carcinogenesis and pH homeostasis: Not only a matter of dysregulated metabolism. *Semin. Canc. Biol.* **43**, 49–65 (2017).
39. Liang, Y. *et al.* Variable effects of maturity-onset-diabetes-of-youth (MODY)-associated glucokinase mutations on substrate interactions and stability of the enzyme. *Biochem. J.* **309**, 167–173 (1995).

Acknowledgements

We thank Maria Angeles Navas (Complutense University of Madrid) for the GST-GCK construct. The study was supported by the Czech Science Foundation project 15-03834Y and Charles University in Prague projects Primus/MED/32, GA UK 1428218 and 260387/SVV/2017.

Author Contributions

D.S. performed the experiments. D.S. and P.H. conceived and designed the experiments, analyzed the data, wrote the paper, are responsible for the integrity of this work, revised the article's intellectual content and approved the final version.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-47883-1>.

Competing Interests: D.S. and P.H. have been funded by the Czech Science Foundation project 15-03834Y and Charles University in Prague projects Primus/MED/32, GA UK 1428218 and 260387/SVV/2017. The authors declare no other competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Miroslav Těšínský, Daniela Šimčíková, Petr Heneberg

First evidence of changes in enzyme kinetics and stability of glucokinase affected by somatic cancer-associated variations

Biochimica et Biophysica Acta – Proteins and Proteomics (2019) 1867: 213-218

Abstract:

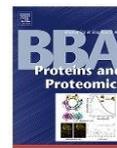
Recent investigation of somatic variations of allosterically regulated proteins in cancer genomes suggested that variations in glucokinase (GCK) might play a role in tumorigenesis. We hypothesized that somatic cancer-associated GCK variations include in part those with activating and/or stabilizing effects. We analyzed the enzyme kinetics and thermostability of recombinant proteins possessing the likely activating variations and the variations present in the connecting loop I and provided the first experimental evidence of the effects of somatic cancer-associated GCK variations. Activating and/or stabilizing variations were common among the analyzed cancer-associated variations, which was in strong contrast to their low frequency among germinal variations. The activating and stabilizing variations displayed focal distribution with respect to the tertiary structure, and were present in the surroundings of the heterotropic allosteric activator site, including but not limited to the connecting loop I and in the active site region subject to extensive rearrangements upon glucose binding. Activating somatic cancer-associated variations induced a reduction of GCK's cooperativity and an increase in the affinity to glucose (a decline in the $S_{0.5}$ values). The hotspot-associated variations, which decreased cooperativity, also increased the half-maximal inhibitory concentrations of the competitive GCK inhibitor, *N*-acetylglucosamine. Concluded, we have provided the first convincing biochemical evidence establishing GCK as a previously unrecognized enzyme that contributes to the reprogramming of energy metabolism in cancer cells. Activating GCK variations substantially increase affinity of GCK to glucose, disrupt the

otherwise characteristic sigmoidal response to glucose and/or prolong the enzyme half-life. This, combined, facilitates glucose phosphorylation, thus supporting glycolysis and associated pathways.



Contents lists available at ScienceDirect

BBA - Proteins and Proteomics

journal homepage: www.elsevier.com/locate/bbapap

First evidence of changes in enzyme kinetics and stability of glucokinase affected by somatic cancer-associated variations



Miroslav Těšínský, Daniela Šimčíková, Petr Heneberg*

Charles University, Third Faculty of Medicine, Prague, Czech Republic

ARTICLE INFO

Keywords:
 Connecting region I
 Hexokinase D
 Lung cancer
 Melanoma
 Persistent hypoglycemic hyperinsulinemia of infancy
 Somatic nonsynonymous substitutions

ABSTRACT

Recent investigation of somatic variations of allosterically regulated proteins in cancer genomes suggested that variations in glucokinase (GCK) might play a role in tumorigenesis. We hypothesized that somatic cancer-associated GCK variations include in part those with activating and/or stabilizing effects. We analyzed the enzyme kinetics and thermostability of recombinant proteins possessing the likely activating variations and the variations present in the connecting loop I and provided the first experimental evidence of the effects of somatic cancer-associated GCK variations. Activating and/or stabilizing variations were common among the analyzed cancer-associated variations, which was in strong contrast to their low frequency among germinal variations. The activating and stabilizing variations displayed focal distribution with respect to the tertiary structure, and were present in the surroundings of the heterotropic allosteric activator site, including but not limited to the connecting loop I and in the active site region subject to extensive rearrangements upon glucose binding. Activating somatic cancer-associated variations induced a reduction of GCK's cooperativity and an increase in the affinity to glucose (a decline in the $S_{0.5}$ values). The hotspot-associated variations, which decreased cooperativity, also increased the half-maximal inhibitory concentrations of the competitive GCK inhibitor, *N*-acetylglucosamine. Concluded, we have provided the first convincing biochemical evidence establishing GCK as a previously unrecognized enzyme that contributes to the reprogramming of energy metabolism in cancer cells. Activating GCK variations substantially increase affinity of GCK to glucose, disrupt the otherwise characteristic sigmoidal response to glucose and/or prolong the enzyme half-life. This, combined, facilitates glucose phosphorylation, thus supporting glycolysis and associated pathways.

1. Introduction

Glucokinase (GCK), also known as hexokinase IV or D, serves as a key glucose sensor in pancreatic β -cells and the liver. In addition to β -cells and hepatocytes, GCK is expressed in α - and δ -cells of the pancreatic islets, enteroendocrine cells of the stomach, L-cells of the intestine, hypothalamus, pituitary, and epithelial cells of the respiratory tract. Combined, these cell types are believed to form a network of glucose sensing cells that maintain systemic glucose homeostasis [1,2]. In contrast to other hexokinases, GCK is characteristic with its low affinity for glucose, apparent cooperativity with glucose and a lack of inhibition by the reaction product, glucose-6-phosphate (G6P). Patients heterozygous for activating variations in the *GCK* gene manifest persistent hypoglycemic hyperinsulinemia of infancy (PHHI), whereas

inhibitory variations induce maturity-onset diabetes of the young (*GCK*-*MODY*) when one allele is affected or permanent neonatal diabetes mellitus (PNDM) when both *GCK* alleles are inactivated. Somatic *GCK* variations are known from cancer tissues, particularly from skin cancer and colorectal carcinoma [3], but their role was never analyzed in detail.

Glucose phosphorylation activity with low affinity for glucose, which is characteristic for GCK, has been observed in a range of cancer cell lines [4]. The GCK activity could be beneficial for cancer cell growth and survival. The GCK-mediated glucose phosphorylation itself is necessary for maintaining the glucose concentration gradient that allows glucose entry into cells. The phosphorylated glucose feeds the glycolytic pathway, glycogenesis, hexosamine and nucleotide biosynthesis, and the pentose phosphate pathway, which allows, among

Abbreviations: BAD, Bcl-2 agonist of cell death; COSMIC, Catalogue of somatic mutations in cancer; G6P, glucose-6-phosphate; GCK, glucokinase; GlcNAc, *N*-acetylglucosamine; GSIR-T, glucose-stimulated insulin release; GST, glutathione-S-transferase; MODY, maturity-onset diabetes of the young; PHHI, persistent hypoglycemic hyperinsulinemia of infancy; PNDM, permanent neonatal diabetes mellitus; RAI, relative activity index

* Corresponding author at: Third Faculty of Medicine, Charles University, Ruská 87, CZ-100 00 Prague, Czech Republic.

E-mail address: petr.heneberg@lf3.cuni.cz (P. Heneberg).

<https://doi.org/10.1016/j.bbapap.2018.12.008>

Received 13 July 2018; Received in revised form 7 December 2018; Accepted 20 December 2018

Available online 24 December 2018

1570-9639/© 2018 Elsevier B.V. All rights reserved.

other things, caspase-2 inactivation and peroxide inactivation. The interaction of GCK with Bcl-2 agonist of cell death (BAD) integrates glycolysis with apoptosis [5–8]. GCK-BAD signaling forms a feedback loop as phosphorylated BAD is required for activation of the mitochondrial-tethered portion of GCK resulting in enhanced glucose and energy metabolism and survival [9–11]. GCK catalyzes the first rate-limiting step of the glycolytic pathway. Thus, activating GCK variations would likely support cancer onset and progression and contribute to the glycolytic phenotype of cancer cells. With two exceptions, all 17 previously identified activating GCK variations are clustered in the surroundings of the heterotropic allosteric activator site and were reported to cause PHHI [12–14]. This site is also targeted by multiple GCK activators [2,15–17]. Both the activating variations (for example p.V91L) and the GCK activators are capable of inducing cellular proliferation [17–20], including the proliferation of cancer cell lines, such as INS-1 [e.g., 18], which supports the role of GCK as a putative proto-oncogene. However, direct evidence for a link between the glucokinase activators and tumor formation or proliferation of GCK-expressing cells is still absent.

Despite direct evidence is absent, the recent investigation of somatic variations of allosterically regulated proteins in cancer genomes has suggested that variations in GCK might play a role in tumorigenesis [21]. The specific activity of GCK is higher than the specific activities of the other hexokinases. We assume that pro-carcinogenic phenotype of GCK variations should likely be associated with retained or increased specific activity and retained lack of inhibition by the reaction product, which is advantageous for unlimited growth. Pro-carcinogenic GCK variations need to increase the GCK affinity to glucose and eliminate the sigmoid kinetics of the GCK response to increasing glucose concentrations, resembling the non-cooperative kinetics that is characteristic for other hexokinases or for GCK after its interaction with activators. Such modified GCK would have a potential to play a pro-carcinogenic role similar to that caused by increased expression of hexokinase 2 and hexokinase 1 in a number of cancers [22–25]. Whether somatic cancer-associated GCK variations have a driver or passenger phenotype and whether they have any role in cancer onset and progression remains enigmatic. We hypothesized that somatic cancer-associated GCK variations include in part those with activating effects on the enzyme kinetics or increasing protein stability. To test this hypothesis, we employed recently proposed GCK-specific settings for computational algorithms [12] that allowed selecting somatic cancer-associated variations, which were most likely to be activating concerning the enzyme kinetics. We analyzed the enzyme kinetics and thermostability of recombinant proteins possessing the likely activating variations and the variations present in the allosteric activator site and provided the first experimental evidence of the effects of somatic cancer-associated GCK variations.

2. Materials and methods

We prepared a series of GCK constructs carrying somatic cancer-associated nonsynonymous substitutions retrieved from the Catalogue of somatic mutations in cancer (COSMIC) database [3]. As of June 5, 2017, 106 cases with 88 nonredundant nonsynonymous substitutions in GCK have been listed in COSMIC v81 database. In the present study, we investigated the effects of 18 somatic missense variations in GCK (Table S1) on enzyme kinetics and the thermostability of the enzyme. To select variations for further enzymological analysis, we applied two independent approaches. First, we used a combination of the prediction algorithms SNAP2 [26] and EVmutation [27] and set the search parameters as recently suggested [12]. The EVmutation scores were unavailable for four of the 88 nonredundant nonsynonymous substitutions; thus, we excluded these four variations from the present study. For further testing, we selected the variations that displayed an EVmutation score > -2.39 , and a SNAP score < -50 , which should allow highly specific selection of variations that are associated with

activating or at least neutral phenotypes. Using the above approach, we selected 11 variations (p.Q24H, p.E27K, p.K104E, p.Q138H, p.H156Q, p.E157K, p.E312K, p.V338L, p.R345H, p.R358P, and p.S433N). Second, we tested six somatic cancer-associated variations (p.R63C, p.R63H, p.S64F, p.T65I, p.P66S, and p.G68S) that are part of the connecting loop I spanning amino acids 64–72 and participating in the binding of allosteric activators [2,28,29] and the somatic cancer-associated variation p.Q106H.

We introduced each of the nonsynonymous substitutions by means of site-directed mutagenesis (primers are disclosed in Table S1) into pGEX-5X-2-GCK, which encoded a glutathione-S-transferase (GST)-tagged wild type GCK isoform 1 (kind gift from M. A. Navas [30]). We expressed and purified the recombinant GST-GCK as described previously [31]. We determined protein concentrations using a Bradford assay (Serva, Heidelberg, Germany) and assessed the purity of the GST-GCK densitometrically based on Coomassie blue-stained 10% SDS-PAGE gels using ImageJ (NIH, Bethesda, MD).

We measured the GCK activity spectrophotometrically using a coupled reaction with glucose-6-phosphate dehydrogenase as described previously [32]. To determine k_{cat} and $S_{0.5}$ we used 11 glucose concentrations (0.1–150 mM) in the presence of 5 mM ATP. To determine $K_{M ATP}$, we used ten ATP concentrations (0.05–10 mM) in the presence of 50 mM glucose. We quantified the effects of the competitive inhibitor *N*-acetylglucosamine (GlcNAc) as the half maximal inhibitory concentrations (IC_{50}), which were measured using eight GlcNAc concentrations (0.1–10 mM) in the presence of 5 mM ATP and 5 mM glucose. We computed the kinetic parameters of the wild type GST-GCK and its somatic cancer-associated variations (η_{11} , $S_{0.5}$, $K_{M ATP}$ and IC_{50} for GlcNAc) as described previously [12]. Based on these parameters, we calculated the threshold for glucose-stimulated insulin release (GSIR-T) and the relative activity index (RAI) as described previously [33].

Furthermore, we measured thermostability of the wild type GST-GCK and its somatic cancer-associated variations at 30 °C, 37 °C, 42 °C and 45 °C in the course of a 100 min incubation at the indicated temperature. All proteins were diluted to 100 $\mu\text{g}\cdot\text{ml}^{-1}$. We measured the GCK activity in the presence of 50 mM glucose and 5 mM ATP.

The results are shown as the mean \pm SEM. Following a Shapiro-Wilk normality test and Levene's equal variance test, data were either analyzed using one-way ANOVA or Kruskal-Wallis ANOVA on ranks. For the post-tests, we used Dunnett's multiple comparison tests. We calculated the Pearson product moment correlation coefficient and the Spearman rank order correlation coefficient in order to correlate the IC_{50} of GlcNAc and the Hill coefficient.

3. Results

3.1. Computational pre-selection of likely activating variations

Using simultaneously two state-of-the-art computational approaches, EVmutation and SNAP2, we identified a series of GCK variations that were likely to cause activating or neutral phenotypes among those reported in COSMIC. We found the most of the cancer-associated variations (53 variations, 63.1%) predicted as likely inhibitory variations, displaying an EVmutation score ≤ -2.39 and a SNAP score ≥ -50 . In contrast, the two computational approaches agreed on the prediction of likely activating or neutral phenotypes for only 11 variations (13.1%). As was shown previously [12], these computational approaches were unable to distinguish between activating and neutral variations. Thus, we further analyzed the whole pool of likely activating or neutral variations together with those present in the connecting loop I.

3.2. Enzymatic characterization

We performed a kinetic analysis with the recombinant wild type

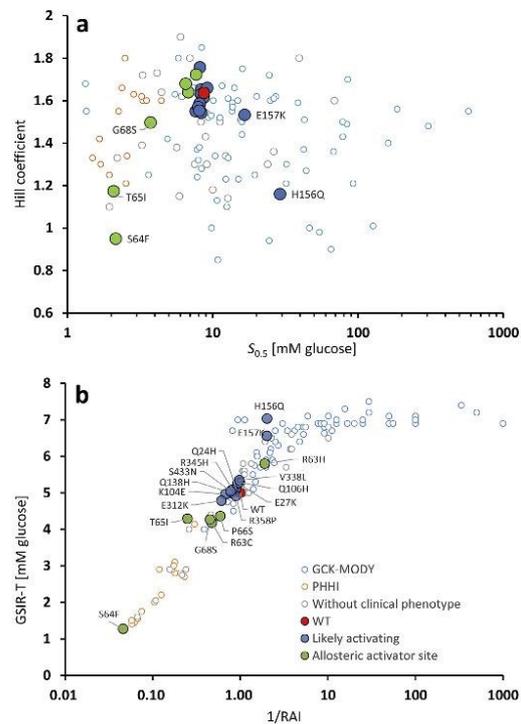


Fig. 1. Effects of somatic cancer-associated GCK variations on enzyme kinetics. (a) Effect of variations on the affinity to glucose ($S_{0.5}$) and cooperativity (n_H). (b) Effects of variations on GSIR-T. To indicate the previously reported extent of both activating and inactivating variation-induced changes in GCK, previously published data on enzyme kinetics of MODY- and PHHI-associated germinal variations [12] are plotted on the background of both subfigures and visualized using open circles.

GST-GCK and its selected variations, for which kinetic constants were not reported yet, except p.T65I, which was already published previously as causing hypoglycemia and thus causing an activating phenotype [16,29,33,34]. All of the variations selected based on the computational algorithms did not change their cooperativity (displayed only negligible changes in their Hill coefficients), except p.H156Q, which nearly lost its cooperativity ($n_H = 1.16 \pm 0.01$). The same set of variations did not display any differences in their affinity for glucose, except, again, p.H156Q and p.E157K, both of which were associated with increases in $S_{0.5}$ ($S_{0.5} = 29.0 \pm 0.7$ and 16.6 ± 0.2 mM of glucose, respectively) (Fig. 1a). The variations p.H156Q and p.V338L decreased significantly their affinity for ATP (a triplication of ATP K_M , Table S2). In contrast, when focusing on the six tested somatic cancer-associated variations, which were selected based on their position in the connecting loop I regardless of their predicted phenotype, three of them (p.S64F, p.T65I and p.G68S) significantly increased the affinity for glucose and/or decreased the Hill coefficient of the affected enzymes (Fig. 1a). The variations in the allosteric activator site did not induce any significant changes in the ATP K_M values (Table S2).

Based on the outcomes of the enzyme kinetics measurements (Table S2), we calculated the GSIR-T and RAI for all of the analyzed variations and compared them to previously published experimental data obtained with variations associated with GCK-MODY and PHHI. We found

that both the GSIR-T and RAI were either insensitive to the variations selected based on the computational algorithms or that a mild to severe increase of the GSIR-T was suggested in the case of the above-mentioned variations p.H156Q and p.E157K (Fig. 1b). In contrast, five of the six variations, which were selected based on their position in the allosteric activator site, caused a decrease in the GSIR-T, and four of them affected the RAI as well. This included the variation p.S64F, which displayed the lowest GSIR-T (1.27) and 1/RAI (0.046) among all of the GCK variations tested in this or any of previous studies, suggesting a strongly activating phenotype (Fig. 1b).

3.3. Thermostability

We measured the activity of wild type GST-GCK and the selected GST-GCK variations at four temperatures ranging from 30 °C to 45 °C to analyze whether these variations affect thermostability of GST-GCK. Under the experimental conditions, the differences between wild type and the tested GST-GCK variations were negligible at 30 °C and 37 °C, but there were prominent differences at 42 °C or 45 °C. In contrast to the results obtained during enzyme kinetics measurements, the thermostability changed predominantly in response to most of the variations selected based on the computational algorithms, whereas changes in the thermostability were negligible in case of variations selected based on their position in or near the allosteric activator site (Fig. 2). Four of the algorithm-selected variations, namely, p.S433N (Fig. 2c), p.K104E, p.R345H and p.Q24H, displayed strong stabilizing effects. Other four algorithm-selected variations, namely, p.R358P (Fig. 2d), p.V338L (Fig. 2e), p.E312K and p.E27K, destabilized the protein and led to decreased or completely undetectable activity following GCK incubation at 45 °C (Fig. 2a).

3.4. Competitive inhibition by GlcNAc

We measured the competitive inhibition of wild type GST-GCK and its selected variations by GlcNAc expressed as the IC_{50} . Under the experimental conditions, the IC_{50} changed in response to two of the six variations selected based on their position in or near the allosteric activator site. In contrast, the IC_{50} remained at values similar to those of wild type GST-GCK in proteins possessing variations selected based on computational algorithms (Fig. 3). Importantly, the IC_{50} of GlcNAc (Fig. 3b) correlated with the Hill coefficient (Fig. 3a) (Pearson -0.777 , $p < 0.001$; Spearman -0.393 , $p = .03$; $n = 31$).

4. Discussion

Previous studies focused exclusively on the effects of germinal GCK variations in patients with impaired glucose homeostasis; their checklists were published by Osbak et al. [35] and later updated by Šimčíková et al. [12]. In the present study, we have provided the first biochemical characterization of a group of somatic cancer-associated variations in GCK. In agreement with the initial hypothesis, we found that a subset of somatic cancer-associated variations have activating effects on the enzyme kinetics and/or increase protein stability. In contrast to the frequency of activating somatic cancer-associated variations, activating germinal GCK variations are infrequent and cause a very rare form of familial hyperinsulinism. The number of previously reported somatic cancer-associated missense GCK variations is low. Those, which were predicted to be inhibitory according to SNAP2 and EVMutation or were identified as neutral or inhibitory in the present study, displayed a random distribution across the GCK coding sequence. Their distribution resembled that of the previously described distribution of GCK-MODY-associated germinal variations in this enzyme [35]. In contrast, the distribution of activating somatic cancer-associated missense GCK variations was focal with respect to the tertiary GCK sequence (Fig. 4).

All the activating somatic cancer-associated GCK variations

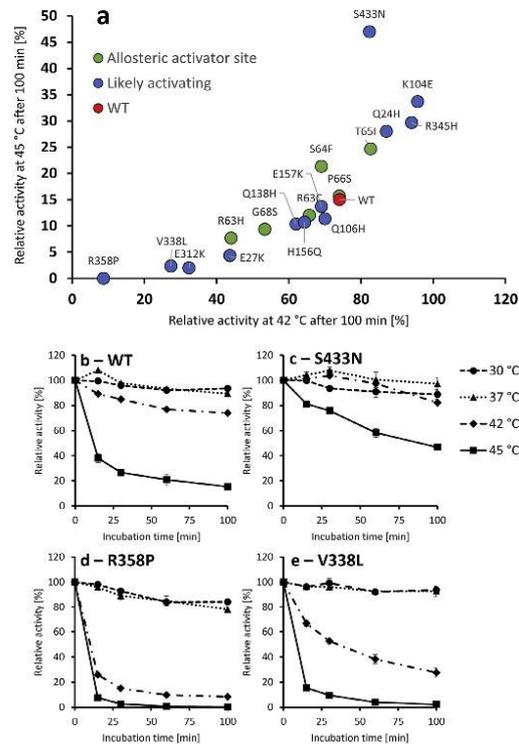


Fig. 2. Effects of somatic cancer-associated GCK variations on thermostability of GCK. (a) Comparison of relative activities of wild type GST-GCK and its somatic cancer-associated variations after 100 min of incubation at 42 °C and 45 °C. (b–e) Kinetics of the thermostability of the wild type (b), p.S433N (c), p.R358P (d) and p.V338L (e) GST-GCK at 30 °C, 37 °C, 42 °C and 45 °C. Relative activities values are shown as the means \pm SEM ($n = 3$).

acquired the ability for non-cooperative binding of glucose and/or simultaneously increased the IC_{50} of GlcNAc. Despite the fact that the activating variations were seemingly interspersed throughout the GCK primary sequence, they were in fact clustered in a tertiary structure region termed the heterotropic allosteric activator site [34] or in its very close proximity (Fig. 4). This clustering of activating somatic cancer-associated variations resembled the more prominent ones in genes with larger numbers of known activating cancer-associated variations, such as *TP53* [36] or *BRAF* [37]. In agreement with Gloyn et al. [34], some of the activating variations were also present in the active site region subject to extensive rearrangements upon glucose binding, which consists of amino acids 151–180. This region is crucial for GCK cooperativity regulation and is the only GCK structure where extensive rearrangements occur upon glucose and ATP binding [29]. Part of this region, namely amino acids 154–164, even oscillates between a structure of mobile loop in unliganded GCK and a β -hairpin in glucose-bound GCK [29]. Variations introduced into this region have a potential to suppress fully GCK cooperativity as confirmed experimentally by Whittington et al. [29]. In the present study, two of the cancer-associated variations that eliminated the cooperativity of the enzyme were present in this region. However, strong inhibition of GCK cooperativity was not limited to the variations in amino acids 151–180, but also was a characteristic feature of activating GCK variations that were located to

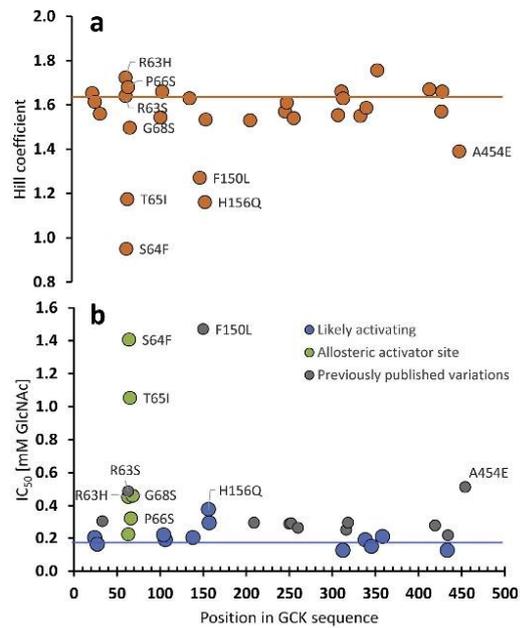


Fig. 3. Effects of somatic cancer-associated GCK variations on cooperativity (a) and inhibition by GlcNAc (b). The variations are sorted according to their position in the GCK primary sequence. To indicate the previously reported extent of both activating and inactivating variation-induced changes in GCK, previously published data on enzyme kinetics of MODY-associated germinal variations [12] are indicated. The mean n_H and IC_{50} of the wild type GST-GCK are indicated by orange and blue solid lines, respectively.

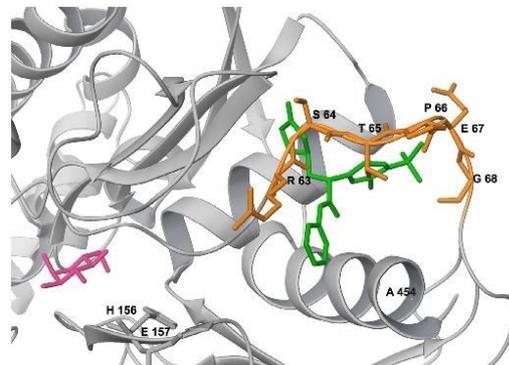


Fig. 4. Structure of the heterotropic allosteric activator site of human glucokinase (PDB ID: 4N07 [51]) with positions of activating somatic cancer-associated GCK variations indicated. All activating variations found in course of the present study, and all but two activating variations published previously as causing hypoglycemia, are clustered in the surroundings of the allosteric activator site and the active site region subject to extensive rearrangements upon glucose binding.

any parts of the heterotropic allosteric activator site and its surroundings (Fig. 3a). GCK cooperativity is directly driven by the equilibrium distribution and conversion of states of the unliganded enzyme, both of

which are impacted by the physiological concentrations of glucose or by the addition of allosteric activators [29]. It remains to be tested, whether structural changes caused by activating variations affect the equilibrium distribution and conversion of GCK states. These effects would simply explain why most somatic cancer-associated variations, which were present in the allosteric activator site and its surroundings, had activating effects on the enzyme kinetics and/or increased protein stability.

The reduction in GCK's cooperativity and a decrease in the glucose $S_{0.5}$ value stand behind the development of familial hyperinsulinism caused by activating GCK variations [34,38,39]. Correspondingly, we found that the somatic cancer-associated variations include those that have lost cooperativity (p.S64F, p.T65I and p.H156Q), including one that was reported as a germinal PHHI-inducing variation (p.T65I), and including one that displayed even stronger declines in GSIR-T and 1/RAI than any activating variations that were reported previously to induce PHHI (p.S64F; Fig. 1b). Correspondingly, the two of the three focally-distributed somatic cancer-associated variations with the decreased Hill coefficients were also decreased in their glucose $S_{0.5}$ values (Fig. 1a). Thus, the properties (Fig. 1) and distribution (Fig. 3) of somatic cancer-associated variations and germinal PHHI-associated variations are similar. These similarities suggest that the surroundings of the allosteric activator site and the active site region subject to extensive rearrangements upon glucose binding serve as hotspots of variations that have a potential to induce familiar hyperinsulinism, and the dysregulation of which is likely to be beneficial for the dysregulated growth of cancer cells. The *in vivo* effects of activating somatic cancer-associated variations found in these hotspots remain to be tested.

We found that most of the somatic cancer-associated variations from the allosteric activator site and some other cancer-associated or previously reported variations both prevented cooperativity and increased the IC_{50} of GlcNAc (Fig. 3). GlcNAc, the glucose analog, serves as a competitive inhibitor of GCK. Increasing GlcNAc concentrations induce a progressive decrease in the Hill coefficient of GCK [40,41]. It has been suggested that the GlcNAc binds to the glucose-binding conformer of GCK, or perhaps also to the GCK-ATP complex that binds glucose only slowly, but not to the free enzyme that binds glucose rapidly. The GlcNAc binding clearly differs from that of palmitoyl-CoA, which also serves as a partial competitive inhibitor of GCK, but which likely binds to a different site and does not change the GCK cooperativity [42–44]. The data obtained in the present study, matched with those obtained under identical laboratory conditions by Šimčíková et al. [12], suggest that the decreased Hill coefficient and increased IC_{50} of GlcNAc can be observed in response to variations in the allosteric activator site. These include variations in or around the electron donor/acceptor interaction site (p.S64 and p.T65) and in another part of the allosteric activator site (p.A454), where an unspecified insertion in the p.A454 was already known to lead to PHHI-associated phenotype [16]. The decreased cooperativity and increased resistance to inhibition by GlcNAc are also associated with variations in the amino acids 151–180 loop, namely, p.H156, which serves as the ATP-binding site [45], and p.F150, which serves as a BAD phospho-mimetic of the Bcl-2 homology 3 alpha-helix binding site [46].

In conclusion, we have provided convincing biochemical evidence establishing GCK as a previously unrecognized enzyme that may contribute to the reprogramming of energy metabolism in cancer cells, one of the hallmarks of cancer [47]. We found that somatic cancer-associated variations in GCK are often activating or stabilizing the enzyme and further experiments should elucidate whether the extent of the contribution of GCK is similar to that reported recently concerning other isoforms of glycolytic enzymes [48,49]. Increased glycolysis stimulates ATP, NADPH and ribose-5-phosphate production, which allow the biosynthesis of lipids and nucleic acids [50]. Activating GCK variations substantially increase the affinity of GCK to glucose, disrupt the otherwise characteristic sigmoidal response to glucose and/or prolong the enzyme half-life. This, combined, facilitates glucose

phosphorylation, thus supporting glycolysis and associated pathways. Future research should address the effects on proliferation of GCK-dependent insulinoma cells, such as the INS-1 cell line [18]. It is also unclear, whether somatic GCK variations play any role in proliferation and survival of cancer cells that express other hexokinases with much higher affinities for glucose, or whether their physiological effects are limited to cells that express GCK only.

Acknowledgements

We thank Maria Angeles Navas (Complutense University of Madrid) for the wild type GCK construct.

Funding

This work was supported by the Czech Science Foundation [grant number 15-03834Y] and Charles University in Prague [grant numbers Primus/MED/32, GA UK 1428218 and 260387/SVV/2017].

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.bbapap.2018.12.008>.

References

- [1] T.L. Jetton, et al., Analysis of upstream glucokinase promoter activity in transgenic mice and identification of glucokinase in rare neuroendocrine cells in the brain and gut, *J. Biol. Chem.* 269 (1994) 3641–3654.
- [2] F.M. Matschinsky, et al., The network of glucokinase-expressing cells in glucose homeostasis and the potential of glucokinase activators for diabetes therapy, *Diabetes* 55 (2006) 1–12.
- [3] S.A. Forbes, et al., COSMIC: somatic cancer genetics at high-resolution, *Nucleic Acids Res.* 45 (2017) D777–D783.
- [4] M. Board, et al., High K_m glucose-phosphorylating (glucokinase) activities in a range of tumor cell lines and inhibition of rates of tumor growth by the specific enzyme inhibitor mannoheptulose, *Cancer Res.* 55 (1995) 3278–3285.
- [5] N.N. Danial, et al., BAD and glucokinase reside in a mitochondrial complex that integrates glycolysis and apoptosis, *Nature* 424 (2003) 952–956.
- [6] N.-L.C. Bui, et al., Bad phosphorylation as a target of inhibition in oncology, *Cancer Lett.* 415 (2018) 177–186.
- [7] N.N. Danial, BAD: undertaker by night, candyman by day, *Oncogene* 27 (2008) S53–S70.
- [8] P. Jiang, et al., The Bad guy cooperates with good cop p53: Bad is transcriptionally up-regulated by p53 and forms a Bad/p53 complex at the mitochondria to induce apoptosis, *Mol. Cell. Biol.* 26 (2006) 9071–9082.
- [9] M.A. Osundiji, et al., BAD modulates counterregulatory responses to hypoglycemia and protective glucoprivic feeding, *PLoS ONE* 6 (2011) e28016.
- [10] A. Giménez-Cassina, et al., Regulation of hepatic energy metabolism and gluconeogenesis by BAD, *Cell Metab.* 19 (2014) 272–284.
- [11] S. Ljubcic, et al., Phospho-BAD BH3 mimicry protects β cells and restores functional β cell mass in diabetes, *Cell Rep.* 10 (2015) 497–504.
- [12] D. Šimčíková, et al., Evidence-based tailoring of bioinformatics approaches to optimize methods that predict the effects of nonsynonymous amino acid substitutions in glucokinase, *Sci. Rep.* 7 (2017) 9499.
- [13] S. Sayed, et al., Extremes of clinical and enzymatic phenotypes in children with hyperinsulinism caused by glucokinase activating mutations, *Diabetes* 58 (2009) 1419–1427.
- [14] N.L. Beer, et al., Discovery of a novel site regulating glucokinase activity following characterization of a new mutation causing hyperinsulinemic hypoglycemia in humans, *J. Biol. Chem.* 286 (2011) 19118–19126.
- [15] J. Grimsby, et al., Allosteric Activators of Glucokinase: potential Role in Diabetes Therapy, *Science* 301 (2003) 370–373.
- [16] F.M. Matschinsky, Assessing the potential of glucokinase activators in diabetes therapy, *Nat. Rev. Drug Discov.* 8 (2009) 399–416.
- [17] A. Nakamura, Y. Terauchi, Present status of clinical deployment of glucokinase activators, *J. Diabetes Investig.* 6 (2015) 124–132.
- [18] Y.S. Oh, et al., Treatment with glucokinase activator, YH-GKA, increases cell proliferation and decreases glucotoxic apoptosis in INS-1 cells, *Eur. J. Pharm. Sci.* 51 (2014) 137–145.
- [19] S. Porat, et al., Control of pancreatic β cell regeneration by glucose metabolism, *Cell Metab.* 13 (2011) 440–449.
- [20] S. Kassem, et al., Large islets, beta-cell proliferation, and a glucokinase mutation, *N. Engl. J. Med.* 362 (2010) 1348–1350.
- [21] Q. Shen, et al., Proteome-scale investigation of protein allosteric regulation perturbed by somatic mutations in 7,000 cancer genomes, *Am. J. Hum. Genet.* 100 (2017) 5–20.
- [22] B. Altenberg, K.O. Greulich, Genes of glycolysis are ubiquitously overexpressed in

- 24 cancer classes, *Genomics* 84 (2004) 1014–1020.
- [23] S.P. Mathupala, et al., Hexokinase II: cancer's double-edged sword acting as both facilitator and gatekeeper of malignancy when bound to mitochondria, *Oncogene* 25 (2006) 4777–4786.
- [24] S.Y. Peng, et al., Aberrant expression of the glycolytic enzymes aldolase B and type II hexokinase in hepatocellular carcinoma are predictive markers for advanced stage, early recurrence and poor prognosis, *Oncol. Rep.* 19 (2008) 1045–1053.
- [25] L.E. Botzler, et al., Hexokinase 2 is a determinant of neuroblastoma metastasis, *Br. J. Cancer* 114 (2016) 759–766.
- [26] M. Hecht, et al., Better prediction of functional effects for sequence variants, *BMC Genomics* 16 (2015) S1.
- [27] T.A. Hopf, et al., Mutation effects predicted from sequence co-variation, *Nat. Biotechnol.* 35 (2017) 128–135.
- [28] J.A. Martínez, et al., Role of connecting loop I in catalysis and allosteric regulation of human glucokinase, *Protein Sci.* 23 (2014) 915–922.
- [29] A.C. Whittington, et al., Dual allosteric activation mechanisms in monomeric human glucokinase, *Proc. Natl. Acad. Sci. U. S. A.* 112 (2015) 11553–11558.
- [30] C.M. García-Herrero, et al., Functional analysis of human glucokinase gene mutations causing MODY2: exploring the regulatory mechanisms of glucokinase activity, *Diabetologia* 50 (2007) 325–333.
- [31] Y. Liang, et al., Variable effects of maturity-onset-diabetes-of-youth (MODY)-associated glucokinase mutations on substrate interactions and stability of the enzyme, *Biochem. J.* 309 (1995) 167–173.
- [32] E.A. Davis, et al., Mutants of glucokinase cause hypoglycaemia- and hyperglycaemia syndromes and their analysis illuminates fundamental quantitative concepts of glucose homeostasis, *Diabetologia* 42 (1999) 1175–1186.
- [33] A.L. Gloyn, et al., F.M. Matschinsky, M.A. Magnuson (Eds.), *Glucokinase and Glycemic Disease: From Basics to Novel Therapeutics*. Frontiers in Diabetes 16, Karger, Basel, 2004, pp. 92–109. Glucokinase and the regulation of blood sugar.
- [34] A.L. Gloyn, et al., Insights into the biochemical and genetic basis of glucokinase activation from naturally occurring hypoglycemia mutations, *Diabetes* 52 (2003) 2433–2440.
- [35] K.K. Osbak, et al., Update on mutations in glucokinase (*GCK*), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia, *Hum. Mutat.* 30 (2009) 1512–1526.
- [36] S. Kato, et al., Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis, *Proc. Natl. Acad. Sci. U. S. A.* 100 (2003) 8424–8429.
- [37] E.R. Cantwell-Dorris, et al., BRAF^{V600E}: implications for carcinogenesis and molecular therapy, *Mol. Cancer Ther.* 10 (2011) 385–394.
- [38] H.B.T. Christesen, et al., Activating glucokinase (*GCK*) mutations as a cause of medically responsive congenital hyperinsulinism: Prevalence in children and characterisation of a novel *GCK* mutation, *Eur. J. Endocrinol.* 159 (2008) 27–34.
- [39] P. Pal, B.G. Miller, Activating mutations in the human glucokinase gene revealed by genetic selection, *Biochemistry* 48 (2009) 814–816.
- [40] M.L. Cárdenas, et al., Suppression of kinetic cooperativity of hexokinase D (glucokinase) by competitive inhibitors. A slow transition model, *Eur. J. Biochem.* 145 (1984) 163–171.
- [41] A. Vandercammen, E. Van Schaftingen, Competitive inhibition of liver glucokinase by its regulatory protein, *Eur. J. Biochem.* 200 (1991) 545–551.
- [42] P.S. Tippet, K.E. Neet, Specific inhibition of glucokinase by long chain acyl coenzymes A below the critical micelle concentration, *J. Biol. Chem.* 257 (1982) 12839–12845.
- [43] P.S. Tippet, K.E. Neet, An allosteric model for the inhibition of glucokinase by long chain acyl coenzyme A, *J. Biol. Chem.* 257 (1982) 12846–12852.
- [44] P.B. Chock, et al., *Enzyme Dynamics and Regulation*, Springer, Dordrecht, 2012.
- [45] Y.N. Kumar, et al., Comparison and correlation of binding mode of ATP in the kinase domains of Hexokinase family, *Bioinformatics* 8 (2012) 543–547.
- [46] B. Szlyk, et al., A phospho-BAD BH3 helix activates glucokinase by a mechanism distinct from that of allosteric activators, *Nat. Struct. Mol. Biol.* 21 (2014) 36–42.
- [47] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, *Cell* 144 (2011) 646–674.
- [48] B. Altenberg, K.O. Greulich, Genes of glycolysis are ubiquitously overexpressed in 24 cancer classes, *Genomics* 84 (2004) 1014–1020.
- [49] P.K. Majumder, et al., mTOR inhibition reverses Akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways, *Nat. Med.* 10 (2004) 594–601.
- [50] S. Ganapathy-Kanniappan, J.F. Geschwind, Tumor glycolysis as a target for cancer therapy: progress and prospects, *Mol. Cancer* 12 (2013) 152.
- [51] P. Petit, et al., The active conformation of human glucokinase is not altered by allosteric activators, *Acta Crystallogr. D Biol. Crystallogr.* 67 (2011) 929–935.

Daniela Šimčíková, Petr Heneberg

**Refinement of evolutionary medicine predictions based on clinical evidence
for the manifestations of Mendelian diseases**

Under revision in *Scientific Reports* (2019)

Abstract:

Prediction methods have become an integral part of biomedical and biotechnological research. However, their clinical interpretations are largely based on biochemical or molecular data, but not clinical data. Here, we focus on improving the reliability and clinical applicability of prediction algorithms. We assembled and curated two large non-overlapping large databases of clinical phenotypes. These phenotypes were caused by missense variations in 44 and 63 genes associated with Mendelian diseases. We used these databases to establish and validate the model, allowing us to improve the predictions obtained from EVmutation, SNAP2 and PoPMuSiC 2.1. The predictions of clinical effects suffered from a lack of specificity, which appears to be the common constraint of all recently used prediction methods, although their predictions are associated with nearly absolute sensitivity. We introduced evidence-based tailoring of the default settings of the prediction methods; this tailoring substantially improved the prediction outcomes. Additionally, the comparisons of the clinically observed and theoretical variations led to the identification of large previously unreported pools of variations that were under negative selection during molecular evolution. The evolutionary variation analysis approach described here is the first to enable the highly specific identification of likely disease-causing missense variations that have not yet been associated with any clinical phenotype.

Title: **Refinement of evolutionary medicine predictions based on clinical evidence for the manifestations of Mendelian diseases**

Authors: **Daniela Šimčíková¹, Petr Heneberg^{1,*}**

Short running title: **Evidence-based predictions of clinical phenotypes**

Authors' affiliation:

¹ Charles University, Third Faculty of Medicine, Prague, Czech Republic

* Corresponding author & the author to whom requests for reprints should be addressed: Petr Heneberg, Third Faculty of Medicine, Charles University, Ruská 87, CZ-100 00 Prague, Czech Republic, Tel: ++420 – 775 311 177, Fax: ++420 – 267 162 710, E-mail: petr.heneberg@lf3.cuni.cz

ABSTRACT

Prediction methods have become an integral part of biomedical and biotechnological research. However, their clinical interpretations are largely based on biochemical or molecular data, but not clinical data. Here, we focus on improving the reliability and clinical applicability of prediction algorithms. We assembled and curated two large non-overlapping large databases of clinical phenotypes. These phenotypes were caused by missense variations in 44 and 63 genes associated with Mendelian diseases. We used these databases to establish and validate the model, allowing us to improve the predictions obtained from EVmutation, SNAP2 and PoPMuSiC 2.1. The predictions of clinical effects suffered from a lack of specificity, which appears to be the common constraint of all recently used prediction methods, although their predictions are associated with nearly absolute sensitivity. We introduced evidence-based tailoring of the default settings of the prediction methods; this tailoring substantially improved the prediction outcomes. Additionally, the comparisons of the clinically observed and theoretical variations led to the identification of large previously unreported pools of variations that were under negative selection during molecular evolution. The evolutionary variation analysis approach described here is the first to enable the highly specific identification of likely disease-causing missense variations that have not yet been associated with any clinical phenotype.

Keywords: computational prediction approaches; missense mutations; monogenic diseases; negative selection; threshold

INTRODUCTION

Computational prediction approaches are an integral part of biomedical and biotechnological research. The prediction algorithms have great potential in precision medicine, particularly with their recent applications in filtering the exome sequencing outcomes for facilitating diagnoses of rare, hardly classifiable, or puzzling disorders suspected of having a genetic origin.¹⁻² The vast majority of coding variations are rare and limited functional data are available.³⁻⁴ This limited availability of evidence-based information is the main argument for the use of prediction algorithms. The prediction algorithms clearly do not outperform evidence-based data in determining the effects of individual variations. However, they allow researchers and clinical geneticists to extrapolate of current knowledge to genes or variations with as yet unknown or uncertain phenotypes. Among the most important modes of use of the prediction algorithms is the assessment of the likely pathogenicity of variations that are discovered *de novo* during exome sequencing studies and in other next-generation sequencing data. Improvements in methods for predicting the pathogenicity of rare coding variations are needed.⁵ Although rare coding variations are often neglected, approximately 100 – 400 of these variations are present in the genome of each human³⁻⁴ and many have been shown to cause inherited diseases.⁶⁻⁷ As we have shown in the pilot study that focused on the glucokinase (GCK), the potential to substantially improve outcomes of already available computational prediction approaches exists when matching them with evidence-based functional data related to clinically reported and/or experimentally analyzed variations in the respective gene.⁸

Most prediction methods assume the *de novo* protein structure and function based on the knowledge of structural features of wild-type proteins and amino acid sequences and their evolutionary conservation.⁹⁻¹¹ Similar approaches have been used to decipher the effects of variations in non-coding sequences.¹² Some approaches, such as PoPMuSiC 2.1¹³, also consider protein thermostability in their estimations.¹⁴⁻¹⁶ The prediction methods may be supervised and thus trained and tested on a properly assembled dataset with reliable annotations.^{15,17} Alternatively, they may be designed as autonomous unsupervised methods, which have better generalization properties and are able to recognize potentially novel types of omics elements,^{12,14,17} but are not resistant to errors incorporated during their development. Most of the prediction methods are based on the evolution-based concept.¹² However, the evolutionary sequence information poorly covers the

additive roles of environmental factors, and the building and interpretation of multiple sequence alignments (MSAs) is still unable to be fully automated.¹⁸⁻¹⁹ Many prediction approaches integrate multiple biophysical characteristics; a classical example of these approaches is SNAP2²⁰. Another strategy that increases the specificity and selectivity is the use of consensus classifiers, such as REVEL⁵, which integrate outcomes of multiple prediction algorithms to eliminate randomly occurring false-positive responses of the individual algorithms. Recently, the traditional approaches were outperformed by an unsupervised prediction method termed EVmutation,¹⁴ which considers epistasis and thus reflects dependencies between positions.²¹⁻²² When the epistasis is reflected in the inference and subsequent use of MSAs, certain variations are labeled as non-acceptable, although they are frequently observed in other positions within the sequence,^{14,23} highlighting the need to incorporate the epistatic approach in individual computational algorithms and consensus classifiers.

In the present study, we hypothesized that the reliability of prediction methods would be improved by switching from *ad hoc* to evidence-based thresholds and provide a proof of concept by modelling and validating this approach for genes associated with Mendelian diseases. We focus on the differences between clinically observed missense variations that are or are not associated with Mendelian diseases and show that the use of evidence-based tailored thresholds substantially improves the prediction of causative disease-associated missense variations (DAVs) among newly identified variations in the course of genomic and proteomic screens.

MATERIALS AND METHODS

We assembled two curated databases of missense variations in genes encoding proteins associated with Mendelian diseases to establish and validate the model (Fig. 1a). When establishing the model, we recognized three categories of variations: 1) “DAVs” represented variations with available evidence of an association with Mendelian diseases. 2) “Partial phenotype-associated” variations were reported to be associated with partial (incompletely manifesting) phenotypes of the same Mendelian diseases. And 3) “No phenotype-associated” variations (NPAVs) were variations with conclusive evidence of the absence of any clinical phenotype associated with their carriers. We predicted the effects of variations using EVmutation¹⁴ based on a specific

epistatic model, SNAP2²⁰, which is based on multiple biophysical characteristics, and PoPMuSiC 2.1¹³ that predicts protein thermostability.

In addition to the clinically observed variations, we calculated and analyzed the predictions for theoretical variations, i.e., variations that have not been clinically observed. We sorted the variations according to a) their localization within/outside protein domains, b) the presence and class of enzymatic activity of the protein, c) the number of nucleotide changes needed to obtain the variation of interest, and d) the American College of Medical Genetics and Genomics (ACMG) classification criteria.²⁴

Selection of genes to establish the model

We selected genes encoding proteins associated with Mendelian diseases according to the availability of a protein structure, inheritance of diseases, and sufficient numbers of clinically observed missense variations (at least nine missense DAVs and at least six missense NPAVs in a region for which the protein structure was available). We retrieved data from the Online Mendelian Inheritance in Man (OMIM; <https://omim.org/>), UniProtKB/Swiss-Prot (<http://www.uniprot.org/>), Protein Data Bank (PDB; <https://www.rcsb.org/>) and Human Gene Mutation Database (HGMD; <http://www.hgmd.cf.ac.uk>). We obtained the evidence for the presence of NPAVs from the ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) and Ensembl (<http://www.ensembl.org/>) databases. We completed information with frequencies of variations and protein domains obtained from the Exome Aggregation Consortium browser (ExAC; <http://exac.broadinstitute.org/>) and the Pfam (<http://pfam.xfam.org/>) database, respectively.

We verified all ambiguous data in the primary literature sources. If we observed conflicting evidence or if conclusive evidence was not available, we removed the variations from the analyses. The factors that led to the removal of variations from the analyzed datasets are listed below. 1) The evidence for only non-Mendelian diseases (e.g., Parkinson disease) was manifested in the carriers of the variation. 2) The variations were listed as benign or likely benign in ClinVar, with high frequencies ($f > 8$) in ExAC, and thus were classified as 1B or higher according to the ACMG criteria for high-quality and abundant data.²⁶ 3) The variations were listed as “DM?” in the HGMD database. These variations denote “a probable/possible pathological mutation, reported to be pathogenic in the corresponding report, but for which (1) the author has indicated that there may be some

degree of uncertainty; (2) the HGMD curators believe greater interpretational caution is warranted; or (3) subsequent evidence has appeared in the literature which has called the putatively deleterious nature of the variant into question".²⁷ 4) Variations for which a disagreement occurred between HGMD (classified as "DM") and ClinVar (classified as "benign" or "likely benign").

We used the key provided in Table 1 to assign of the clinically observed variations. We selected all clinically observed variations, which we used to set the thresholds, using the key described above. Additionally, we included the GCK variations resulting from the systematic literature review provided in 2017 by Šimčíková *et al.*⁸ We classified nine variations as NPAVs based on the recent literature.²⁸⁻³³ We included the hemoglobin variations, which were classified as likely non-phenotypic in the HGMD database, in the NPAVs.

The variations classified in ClinVar as VUS (n = 404) were subjected to the analysis using EVmutation, and SNAP2 scores shifted slightly but significantly towards their pathogenicity compared to the variations classified as benign or likely benign (n = 1589): EVmutation mean \pm SD -4.21 ± 2.54 vs -3.84 ± 2.38 , *t*-test *p* = 0.003; SNAP2 mean \pm SD -0.9 ± 57.34 vs -12.57 ± 55.85 , *t*-test *p* < 0.001. Based on these calculations, we excluded the hemoglobin variations that were classified as likely non-phenotypic in HGMD (n = 100). These variations received EVmutation scores, but not SNAP2 scores, similar to VUS (EVmutation mean \pm SD -4.26 ± 2.25 , *t*-test vs VUS *p* > 0.05; SNAP2 mean \pm SD 15.58 ± 42.69 , *t*-test vs VUS *p* = 0.003).

All variations included in the dataset we used to establish the model were classified according to the ACMG criteria,²⁴ differentiating between those classified as benign (1B, 3B and 5B) and pathogenic (0.5P and 1P). The frequencies of the variations according to the ACMG classification are provided in Table S10.

We retrieved clinical information on 7178 missense variations (Fig. 1a) located within the coding sequences of 44 genes that, if mutated, cause Mendelian diseases. The following genes were included in the dataset we used to validate the model: *AR*, *ATP7A*, *BMP2R*, *BTK*, *CD40LG*, *CDKL5*, *CPOX*, *CYBB*, *DCX*, *DMD*, *EDA*, *ELANE*, *F9*, *FHL1*, *FLNA*, *G6PD*, *GCK*, *GCH1*, *GLA*, *HBB*, *HDAC8*, *HMBS*, *HNF4A*, *HPRT1*, *HSPB1*, *IDS*, *IL2RG*, *ITGA2B*, *KIT*, *MECP2*, *MSH2*, *OTC*, *PDHA1*, *PROC*, *PTEN*, *PTPN11*, *RET*, *SERPING1*, *SH2D1A*, *STK11*, *TGFBR2*, *TP63*, *TTR* and *UROD*. All the analyzed missense variations were limited to those parts of the genes for which structural information was available. We designated 4546 variations as "DAVs", because the evidence for their associations with Mendelian

diseases was available. We designated another 291 variations as “partial phenotype-associated”, because the evidence for their association with partial (incompletely manifesting) phenotypes of the same Mendelian diseases was available. We designated 2093 missense variations as “NPAVs”, because conclusive evidence of the absence of any clinical phenotype associated with their carriers was available. We removed 248 (3.5%) missense variations from the analyses due to inconsistent, insufficient or anomalous data on the phenotypes reportedly associated with these variations. Data reliability in databases appears to be a challenge to the construction of the dataset. Standardized forms of annotations do not currently exist. Additionally, submission processes differ among the databases, ranging from individual to bulk submissions, and are rarely checked for consistency with previously published peer-reviewed studies.³² Therefore, the construction of the comprehensive dataset also prevented or considerably decreased the risk of biases that might arise from errors of omission and commission in databases.

Selection of genes to validate the model

We established the validation dataset consisting of 1723 variations in 63 additional genes associated with autosomal dominant or autosomal recessive diseases to validate the newly reported approach on an independent set of proteins that are associated with Mendelian diseases (Table S8). These 63 genes were not included in the dataset that was used to establish the model. We populated the dataset based on the classifications of variations retrieved from ClinVar. We also verified the allele counts in the ExAC browser, but this information was only available for a limited number of variations in this dataset. Thus, this information was not used in the analyses. The genes included in the dataset that was used to validate the model were: *AARS*, *ABCC6*, *ALDH18A1*, *ARSB*, *AVP*, *CASR*, *CFTR*, *CLCN1*, *CLCN7*, *COL7A1*, *DNM2*, *DSP*, *DYNC1H1*, *ELOVL4*, *FBN1*, *FGF23*, *FGFR3*, *GALNS*, *GBA*, *GJB2*, *GJA3*, *GLB1*, *GNE*, *GUCY2D*, *GUSB*, *HEXA*, *HGSNAT*, *IMPDH1*, *KCNA1*, *LMNA*, *LMNB1*, *LRP5*, *MARS*, *MPZ*, *MYH14*, *MYH3*, *MYH7*, *MYH9*, *MYO6*, *NAGLU*, *NOTCH3*, *NR3C2*, *OPA1*, *PGFRB*, *PKD1*, *PKD2*, *POLG2*, *PRKCG*, *PRPF8*, *RAF1*, *RYR1*, *SGSH*, *SLC4A1*, *SMPD1*, *SOS1*, *SOS2*, *SPAST*, *STAT1*, *STAT3*, *TECTA*, *TERT*, *VCP* and *YARS*. The dataset was composed of the following numbers of variations: 33 benign, 53 benign / likely benign variations, 58 likely benign variations, 475 likely pathogenic variations, 104 likely pathogenic / pathogenic variations and 1000 pathogenic variations (Table S8).

Prediction analyses

For all selected proteins, we employed three methods with distinct approaches and bases. First, we used the unsupervised epistatic model EVmutation¹⁴ with the arbitrary threshold set to zero. Second, we used the supervised method SNAP2²⁰, which is based on multiple biophysical characteristics and trained on annotated databases of clinically observed and/or experimentally tested variations from annotated databases (OMIM, PMD and Swiss-Prot). Third, we used PoPMuSiC 2.1¹³, which predicts protein thermostability. The arbitrary threshold of the EVmutation method was set to zero based on the claim by Hopf *et al.*¹⁴ that “values of ΔE above 0 correspond to more probable mutant sequences (putatively beneficial), values below 0 to less probable mutant sequences (putatively deleterious) and values equal to 0 to equally probable sequences (putatively neutral).” Thus, the arbitrary threshold allowed us to differentiate between the “putatively deleterious” and “putatively beneficial” mutations. Based on these criteria, the variation effect scores were also set to zero for all examined wild-type protein sequences in the protein matrices that were precomputed by Hopf *et al.*¹⁴ (available at <https://marks.hms.harvard.edu/evmutation/>, accessed March 8, 2018). Due to the nature of the EVmutation method, almost no “putatively neutral” variations with a zero EVmutation score were observed, except for the wild-type alleles. Hopf *et al.* applied these settings to changes occurring at the protein level, but predictions of the changes at the level of the whole organism are more challenging.

We used the pre-computed predictions from EVmutation that were listed according to the UniProtKB/Swiss-Prot accession numbers. We computed the predicted effects of amino acid changes identified using SNAP2 according to the NCBI code belonging to relevant protein isoforms. We selected the protein structures with a resolution lower than 2.7 Å (except *GCH1* and *PROC*) and used their PDB codes in the prediction computations employing PoPMuSiC 2.1. In addition to the clinically confirmed variations, we calculated and analyzed the predictions for theoretical variations, i.e., variations that were not clinically observed. We performed these calculations for the protein regions identical to those, we used to analyze the clinically observed variations. We sorted the variations according to a) their localization within/outside of protein domains, b) the presence and class of enzymatic activity of the protein, and c) the number of nucleotide changes needed to obtain the variation of interest. When sorting the variations according to the latter criterion, we split theoretical variations into impossible (157,639 variations) and possible variations (63,698 variations) according to the method reported by Bromberg *et al.*¹⁵

They defined “impossible” amino acid variations as those that require a change of two or three nucleotides in the codon, whereas “possible” variations were defined as amino acids variations that require a change in only a single nucleotide.

GV approach

Many variations that were previously associated with Mendelian diseases have been re-assessed and re-classified as VUSs.³³⁻³⁵ In the present study, we limited the MSAs based on the paradigm of the VUS²⁶ classification, which differentiates VUSs from likely benign variations by analyzing their conservation in other mammalian species. According to multiple indices, the predictions of the effects of the analyzed variations may be improved by implementing MSA analyses. The MSA analyses assume that variations identified in related species are likely neutral (non-pathogenic), whereas variations identified in conserved parts of the amino acid sequence are likely pathogenic. A consensus regarding the inclusion criteria for the analyzed sequences has not been reached. Some authors compare the sequences of all proteins in the respective protein family, while others limit the analyzed sequences to those that are similar to human sequences.³⁴⁻³⁵

We used the GV approach to analyze the MSAs of amino acid sequences of the examined human proteins and their mammalian orthologs.²⁵ The GV approach quantifies the variability in each tested amino acid based on the MSA provided. This approach allowed us to classify the variations into those with GV scores of zero (conserved among mammals) and those with higher GV scores (with at least two sequence variations present in the analyzed MSAs). We assembled the MSAs by implementing the paradigm associated with variants of uncertain significance (VUS), which claims that the variations are considered VUSs if an amino acid residue that is conserved in the corresponding protein in other mammals is altered.²⁶ Thus, for each analyzed protein, we prepared the MSA that contained amino acid sequences of ten mammalian orthologs of the respective gene. Typically, we included a dominant human isoform of the respective protein and complemented it with the corresponding isoform reported from two species of primates (Primates) and one sequence each from carnivores (Carnivora), bats (Chiroptera), rodents (Rodentia), even-toed ungulates or cetaceans (Cetartiodactyla) and insectivorous mammals (Eulipotyphla, which is still listed as Insectivora in the NCBI Nucleotide database). The remaining two orthologs were both represented by marsupials (Metatheria) or by one marsupial and one monotreme

(Monotremata), avoiding monotreme sequences when high-quality reads were not available in the NCBI GenBank database. We retrieved all sequences from the NCBI GenBank database between May 30 and June 4, 2017.

Additionally, we tested two representative genes, *AR* and *PTEN*, to determine whether the addition of more evolutionarily distant sequences and the resulting increase in variability led to an improved correspondence of GV scores with disease associations of analyzed variations. We used the maximum likelihood method to estimate evolutionary divergence in amino acid sequences predicted to be encoded by *AR* and *PTEN* among selected taxonomic groups. For *AR*, we tested 29 amino acid sequences of *AR* orthologs, including the orthologs from ten mammalian species, as specified above. The more evolutionarily distant orthologs included sequences from Testudines (three species), Amphibia (three species), Crocodylia (two species), Squamata (four species), Aves (three species), Euteleostomi (three species) and Chondrichthyes (one species). The NCBI Blast search did not retrieve orthologs that would be homologous with *AR* from more evolutionarily distant species. The *PTEN* protein is more evolutionarily conserved, which allowed us to include more distant taxa. The resulting dataset comprised 31 orthologs, ten of which were from the mammalian species listed above, and others consisted of orthologs from the following taxa: Aves (three species), Squamata (three species), Archelosauria (three species), Teleostei (three species), Chondrichthyes, Coelacanthiformes, Amphibia, Brachipoda, Gastropoda, Mollusca, Echinozoa, Arachnida and Insecta (one species each). We retrieved these sequences from the NCBI GenBank database between October 8 and October 14, 2017. We aligned the amino acid sequences using ClustalW (gap opening penalty of 5 and gap extension penalty of 0.1 for pairwise alignments, gap extension penalty of 0.2 for multiple alignments, and gap separation distance of 4). We manually corrected the alignments for any inconsistencies and replaced shorter sequences with more appropriate sequences. We used only sequences of identical lengths for further analyses. We used the resulting MSAs to calculate the GV scores. For the *AR* and *PTEN* alignments, we performed maximum likelihood fits of the 48 amino acid substitution models, excluding positions containing gaps. For each model, we calculated the Bayesian information criterion, corrected Akaike information criterion and maximum likelihood values. For *AR*, we analyzed 29 sequences with 380 positions in the final dataset. For *PTEN*, we analyzed 31 sequences with 342 positions in the final dataset. We used best-fit models for the subsequent phylogenetic analyses and evolutionary divergence calculations. When building the trees, we

constructed the initial tree using a neighbor-joining algorithm. We built the trees based on both AR and PTEN sequences using the Jones-Taylor-Thornton model. We modeled the non-uniformity of evolutionary rates among sites using a discrete Gamma distribution (+G) with five rate categories. We applied a bootstrapping procedure with 1,000 replicates. We used the maximum likelihood method to estimate evolutionary divergence in the amino acid sequences of AR and PTEN orthologs among selected taxonomic groups. We calculated the number of base differences per site by averaging all sequence pairs between groups (distance) \pm SE and employed a bootstrapping procedure with 1,000 replicates. The models used to estimate inter- and intrasite evolutionary divergence were identical to the models used to construct the respective trees.

REVEL

We calculated the sensitivity and specificity of the predictions retrieved from REVEL to test whether the issue of low specificity is associated with the outcomes of individual computational algorithms or whether it also affects the data obtained using state-of-the-art consensus classifiers.⁵ We used REVEL to test a subset of 21 genes from the dataset that was used to establish the model: *GCK*, *AR*, *PTEN*, *CYBB*, *HNF4A*, *HBB*, *MECP2*, *HDAC8*, *RET*, *PTPN11*, *HPRT1*, *CD40LG*, *CDKL5*, *CPOX*, *DCX*, *DMD*, *EDA*, *UROD*, *TTR*, *FLNA* and *HSPB1*. We provided REVEL scores for 2721 variations, of which 1570 were DAVs, 241 manifested partial phenotypes, and 910 were NPAVs. For the aforementioned genes, we tested the identical set of variations as used to establish the model, except for PTEN p.P103Q, PTEN p.A137F, and four GCK variations, representing amino acid substitutions caused by substitutions of two or three nucleotides. We obtained the REVEL scores from the pre-computed database of REVEL scores that are available for all missense variations retrieved from dbNSFP v2.7, as provided by the authors of REVEL.⁵

Statistical analyses

We calculated the evidence-based thresholds as medians \pm 2 \times SD, which should encompass approximately 95% of the pool of variations used to calculate the threshold. We calculated two types of these thresholds. The sensitivity threshold (true positive rate) was calculated based on the 95% chance of confirming the association of a tested theoretical variation with the respective disease based on the distribution of prediction scores for known DAVs. The specificity threshold (true negative rate) was calculated based on the 95% chance of confirming

the absence of an association of a tested theoretical variation with the respective disease based on the distribution of prediction scores for known NPAVs.

We calculated the weighted means of the scores resulting from the tested prediction methods by assigning each predictor a weight ranging from -100 to +100, where 0 was a threshold and 100 was the maximum value observed within the respective dataset (EVmutation range -12.933 – 3.8104, SNAP2 range -98 – 99, and PoPMuSiC 2.1 range -1.90 – 5.64), and by averaging the values obtained from each of the prediction methods.

We tested the differences between predictions between DAVs and NPAVs, and for domain-associated and other amino acids using a one-tailed *t*-test. Differences in the numbers of DAVs and NPAVs in individual domains were determined using one-tailed *t*-tests with Bonferroni's correction. We tested the differences between variations associated with particular classes of enzymes and proteins without enzymatic functions, and between categories of possible and impossible theoretical variations using the Kruskal-Wallis one-way ANOVA on ranks with Dunn's post-tests (the Kolmogorov-Smirnov normality test yielded $p > 0.05$ for each comparison). We analyzed the difference in the frequency of DAVs and NPAVs among possible and impossible theoretical variations using the χ^2 test, with the number of possible variations normalized to the number of impossible variations. We assessed the differences between DAVs (including multiple phenotypes alone), partial phenotype-associated and NPAVs using the Kolmogorov-Smirnov normality test followed by one-way ANOVA with Tukey's post-tests or Kruskal-Wallis one-way ANOVA on ranks with Dunn's post-tests when the normality tests failed. We did not evaluate phenotypes with less than five associated variations. The data are shown as means \pm SD, unless indicated otherwise. We performed all calculations using SigmaPlot 12.0, and conducted phylogenetic analyses using MEGA 5.2.

RESULTS AND DISCUSSION

Outputs of the calculation of thresholds

We hypothesized that the thresholds of predictions obtained using SNAP2 and PoPMuSiC 2.1 are subject to evidence-based adjustment, similar to the EVmutation threshold. The 95% sensitivity of SNAP2 was ensured by

establishing a general evidence-based threshold at a level of median - 2SD, i.e., $61 - 2 \times 46.51 = -32.02$. However, the use of this threshold increases the percentage of false-positive phenotype predictions from 46% to 79%, which is not acceptable. Similarly, a sensitivity of 95% for PoPMuSiC 2.1 predictions was ensured by establishing a general evidence-based threshold at a level of median - 2SD, i.e., $1.17 - 2 \times 1.08 = -1.00$. However, the use of this threshold increases the percentage of false-positive phenotype predictions from 88% to 99.9%, which is not acceptable. When we combined the three prediction methods, they displayed high sensitivity but low specificity when using both the arbitrary and general evidence-based thresholds.

The absence of any agreement in the predictions of NPAVs and the existence of 58% (arbitrary thresholds) or 45% (general evidence-based thresholds) variations, which were predicted differently using the three methods, was alarming and required a more thorough adjustment of the thresholds to produce reliable prediction outcomes. Thus, we tested the application of weighted means. The application of weighted means did not exert any substantial effect on the sensitivity (92% with arbitrary thresholds or 94% with general evidence-based thresholds) but it decreased the specificity to 39% (arbitrary thresholds) and 31% (general evidence-based thresholds).

This issue would potentially be overcome by applying gene-specific evidence-based thresholds, i.e., the thresholds that were calculated individually for each analyzed gene. However, this approach did not overcome the specificity issue, as the problem associated with the incorrect detection of NPAVs remained. PoPMuSiC 2.1 was more problematic in this regard, as its predictions were so variable and skewed that the threshold set as a mean - 2SD of DAVs often exceeded the range of predictions of NPAVs. Using this approach, PoPMuSiC 2.1 incorrectly detected 515 (24.6%) of NPAVs as associated with an effect, although the other two predictors generated correct predictions for this pool of variations. Thus, the agreement of the three methods on the non-pathogenicity of NPAVs was reached for only five of the 2093 (0.0%) NPAVs.

Next, we tested whether the implementation of two gene-specific evidence-based thresholds per predictor for each gene would be the solution. One threshold was set to 95% sensitivity (i.e., the threshold used above) and the other threshold was set to 95% specificity. When we implemented the new combination of thresholds, the three prediction methods only agreed on the predictions for the effects of 303 variations. Among these

variations, 301 variations (99.3%) were DAVs and two variations (0.7%) were NPAVs. Similar to the previous approach, the problematic outcome was primarily caused by the inclusion of hypervariable predictions generated by PoPMuSiC 2.1. When we excluded PoPMuSiC 2.1 from the analyses, the gene-specific 95% specificity threshold was passed by 763 variations (11.5%), of which 752 variations (98.6%) were DAVs and 11 variations (1.4%) were NPAVs. The gene-specific 95% sensitivity threshold was passed by 622 variations (9.4%), of which 102 variations (16.4%) were DAVs and 520 variations (83.6%) were NPAVs. Thus, these findings provide proof of concept that the evidence-based adjustment of thresholds for EVmutation and SNAP2 enables the highly specific selection of both DAVs and NPAVs. To our knowledge, this approach is the first to allow the highly specific selection of variations that are not associated with any clinical phenotype. Within the tested dataset, the predictable variations accounted for 21% of the tested variations. The other variations were divided into the following three categories: a) the predictions of EVmutation and SNAP2 were contradictory (0.2%), b) one of the two predictors did not exceed either of the two thresholds (30.4%), and c) both predictors did not exceed their thresholds (48.7%). The use of weighted means combined with the two gene-specific evidence-based thresholds per predictor did not improve the outcomes and resulted in 33.5% sensitivity and 93.7% specificity.

When we analyzed the EVmutation outputs alone using the identical two gene-specific evidence-based thresholds per predictor for each gene, the gene-specific 95% specificity threshold was passed by 1236 (18.6%) variations, of which 1188 (96.1%) were DAVs and 48 (3.9%) were NPAVs. The gene-specific 95% sensitivity threshold was passed by 807 (12.2%) variations, of which 164 (20.3%) were DAVs and 643 (79.7%) were NPAVs. Thus, the use of EVmutation alone was associated with a slightly greater number of both false negative and false positive predictions, but provided a prediction for a larger percentage of the analyzed variations. Within the tested dataset, the predictable variations accounted for 31% of the total number of tested variations.

When we analyzed the SNAP2 outputs alone using the identical two gene-specific evidence-based thresholds per predictor for each gene, the gene-specific 95% specificity threshold was passed by 1390 (20.9%) of variations, of which 1343 (96.6%) were DAVs and 47 (3.4%) were NPAVs. The gene-specific 95% sensitivity threshold was passed by 1365 (20.6%) variations, of which 403 (29.5%) were DAVs and 962 (70.5%) were NPAVs. Thus, the use of SNAP2 alone was associated with a slightly greater number of both false negative and false positive predictions but provided a prediction for a larger percentage of the analyzed variations compared to its combination with

EVmutation or to EVmutation alone. Within the tested dataset, the predictable variations accounted for 41 % of the tested variations.

EVmutation under default settings

The arbitrary threshold used for the EVmutation analysis enables the correct prediction of a phenotype for 99.5% of DAVs and 99.7% of partial phenotype-associated variations; this sensitivity is consistent with previously reported data.¹⁴ However, 94.8% of NPAVs were in the same category and were predicted to exert an effect. Thus, the arbitrary zero threshold was associated with only a 5.2% specificity for clinically manifested phenotypes (Fig. 1b).

A high number of false positives was observed for all 44 analyzed genes (Fig. 1c). The EVmutation analysis provided the correct predictions of DAVs for all tested genes (median sensitivity of 100%, minimum sensitivity of 92.3% (*RET*)), but only correctly predicted a negligible fraction of NPAVs (median specificity of 4.4%, minimum specificity of 0% (12 genes), maximum specificity of 20% (*CD40LG*)).

Tailored EVmutation thresholds

The arbitrary threshold does not provide a reliable prediction of the disease association of variations in tested genes. Therefore, we focused on whether the thresholds can be tailored either in a general or gene-specific manner. The median \pm SD of predictions obtained using EVmutation for DAVs reached -6.58 ± 2.23 , whereas the values for NPAVs only reached -3.86 ± 2.41 . Thus, these two groups of variations were not separated to an extent that was sufficient for distinguishing between them based on, for example, their confidence intervals. Nevertheless, when focusing on the gene-specific level, the median values of predictions of the DAVs for any gene were lower than the median values of the predictions of NPAVs within the same genes. The scores and resolution varied across the analyzed genes (Fig. S1a). A sensitivity of 95% was assumed by setting the threshold to the median + 2SD of the DAVs, i.e., $-6.57 + 2 \times 2.22 = -2.13$. Thus, the EVmutation score of -2.13 was considered a general evidence-based threshold. Its use increases the specificity to 21.5%, which is, however, still far from any reliable use of this approach.

Constraints in VUS criteria

The VUS classification differentiates VUSs from likely benign variations based on evidence of their conservation in other mammalian species. We identified the conserved variations with the zero GV scores, i.e., variations that were conserved across the whole class of mammals, including marsupials and/or monotremes. The conserved variations represented 69.7% of NPAVs and 86.2% of DAVs. The conserved variations were associated with slightly lower EVmutation scores for both DAVs and NPAVs (Fig. 1d) compared to variations that affected evolutionarily variable sites. Nevertheless, the EVmutation scores of the four groups of variations overlapped and required further stratification. Thus, we examined the relative proportion of variations with a GV score > 0 individually in each of the 44 analyzed genes (Fig. S2a). All variations in some genes displayed a zero GV score (*AR* and *PTEN*), whereas variations in other genes were poorly conserved (*ELANE*, *PROC* and *CD40LG*). Based on this finding, the arbitrary criteria for the inclusion of protein sequences in the MSAs derived from the VUS criteria were not functional since they did not reflect differences in the conservation of individual genes. Absolute values of the GV scores (degree of conservation of the respective amino acid) were not associated with any differences in clinical phenotypes (Fig. S2a) or EVmutation scores (Fig. S2b) for variations of these amino acids. However, the binary response (zero GV score vs any higher GV score) predicted the stratification of variations into DAVs and NPAVs.

We postulated that the MSAs, which were based on VUS inclusion criteria, were insufficient for the analyses of highly conserved genes, such as *AR* or *PTEN*, because these genes displayed low amino acid sequence divergence among their mammalian orthologs. The solutions consisted of the addition of more evolutionarily distant taxa into the alignments (Figs. 2a and S3). This addition increases the divergence between the analyzed groups of organisms (Tables S1-S2), which is sufficient to generate a pool of informative amino acids that are susceptible to variations during the course of evolution. Although the VUS-based GV score (i.e., the score that was based solely on sequences of mammalian orthologs) did not discriminate between the DAVs and NPAVs, the GV score based on extended MSAs led to a clear differentiation between DAVs and NPAVs. The DAVs were associated with 60 – 80% of amino acids with a zero GV score. In contrast, the NPAVs reached zero scores in 20 – 30% of cases (Fig. 2b-c). Thus, the MSAs used to calculate the GV scores of highly conserved proteins were improved by including sequences from evolutionarily distant organisms until an experimentally or arbitrarily set value of sequence divergence between analyzed groups (≥ 0.1 substitutions per amino acid) was achieved.

Even using these improved settings, a large group of variations were considered DAVs, despite displaying high GV scores (Fig. 2b-c).

Combination of EVmutation with methods based on different approaches

We next focused on improving EVmutation-based predictions by combining them with other state-of-the-art prediction methods that provide numerical outcomes and thresholds, which can easily undergo evidence-based adjustment. Similar to EVmutation, the arbitrary settings of SNAP2²⁰ and PoPMuSiC 2.1¹³ do not correspond to the division of clinically observed variations into DAVs and NPAVs (Fig. S4a-b). For SNAP2, 84% of predictions of DAVs and 54% of predictions of NPAVs were correct. Thus, the percentage of true disease predictions was slightly lower than with EVmutation, but the percentage of true no phenotype predictions was higher by an order of magnitude than with EVmutation. For PoPMuSiC 2.1, we obtained correct predictions for 94% of DAVs and only 12% of NPAVs. Thus, the number of true disease predictions was slightly lower than with EVmutation, and the percentage of true no phenotype predictions was similar to EVmutation. In contrast to EVmutation, the latter two prediction methods were associated with a high variability of predictions between the analyzed proteins (Fig. S4c-d).

We hypothesized that the thresholds of predictions obtained using SNAP2 and PoPMuSiC 2.1 could benefit from being subjected to evidence-based adjustment, similar to the adjustment of the EVmutation threshold. We tested several approaches for calculating the thresholds (see the chapter Outputs of the calculation of thresholds for a detailed description of the applied approaches), but most of these approaches only provided minor or no improvements. Additionally, the PoPMuSiC 2.1 scores were associated with such high overlap of the distribution of DAVs and NPAVs that the outcomes of this method were uninformative. Therefore, we excluded PoPMuSiC 2.1 from further analyses. The approach that led to a substantial improvement in the credibility of predictions was the implementation of two gene-specific evidence-based thresholds per predictor for each gene. One gene-specific threshold was set to 95% sensitivity (i.e., the threshold used above) and the other threshold was set to 95% specificity. For the combination of EVmutation and SNAP2, the predictable variations represented 21% of the total number of tested variations. The predictions were associated with a 98.6% specificity and 83.6% sensitivity. Thus, this result serves as proof of concept that the evidence-based

adjustment of thresholds for EVmutation and SNAP2 enables the highly specific selection of both DAVs and NPAVs. To our knowledge, this approach is the first to enable the highly specific selection of variations that are not associated with any clinical phenotype.

When the two predictors were used alone, the percentage of predictable variations increased (to 31% using EVmutation and 41% using SNAP2), but the specificity and sensitivity decreased. For EVmutation, the specificity was 96.1% and sensitivity was 79.7%. For SNAP2, the specificity was 96.6% and sensitivity was 70.5%. Thus, the use of EVmutation or SNAP2 alone was associated with a slightly higher number of both false negative and false positive predictions but provided a prediction for a larger percentage of the analyzed variations when compared to their combination.

Factors contributing to the variability within the analyzed dataset

The predictions of the effects of DAVs and NPAVs differed for variations located within or outside of the protein domains (*t*-test $p < 0.001$ each, for EVmutation and SNAP2, respectively). The predictions of the effects of DAVs differed for variations located within and outside of the protein domains (*t*-test $p < 0.001$ each, for EVmutation and SNAP2, respectively). In contrast, the NPAVs did not display any significant difference between their pools located within and outside of the protein domains (*t*-test $p > 0.05$ each, for EVmutation and SNAP2, respectively) (Fig. 2d). Thus, the predictions of the variations present within protein domains displayed a higher amplitude (EVmutation -2.722 vs -1.973, and SNAP2 71 vs 47). When focusing on particular domain types, the differences between DAVs and NPAVs were significant for all major domain types (*t*-test with the Bonferroni's correction $p < 0.001$), except the globin domain (*t*-test with the Bonferroni's correction $p > 0.05$ for both predictors) and ligand-binding domain of nuclear hormone receptor (SNAP2 *t*-test with the Bonferroni's correction $p > 0.05$) (Fig. 2e and Table S3). In the combination approach, the variations that were located within catalytically active protein domains (e.g., tyrosine kinases or serine-threonine kinases) were easier to predict than variations that were located outside of any domains. The prediction of variations located within certain protein domains lacking intrinsic enzymatic activity was highly problematic, but certain enzymatically inactive domains (e.g., the SH2 domain) were still associated with an acceptable resolution of the predictions. The rigidity of the SH2 domain structure (needed for pTyr binding)³⁶ was likely responsible for this difference in

prediction outcomes compared with the globin domains. The globin domains maintain their function, regardless of their low sequence identity, as long as the hydrophobic core and hydrophilic surface are maintained.³⁷ The predictions of variations in the amino acid sequences of enzymes also showed a better resolution than those of variations located in proteins without enzymatic functions (Fig. 2f). Only differences between the DAVs (but not NPAVs) of proteins without enzymatic function and any of the four enzyme classes tested were significant (Kruskal-Wallis one-way ANOVA on ranks with Dunn's post-tests $p < 0.05$ each; Table S4). Future algorithms should match the predictions with protein attributes, such as the presence of specific protein domains.³⁸ The binary presence/absence information for the location in protein domains is used to identify driver and passenger somatic mutations involved in oncogenesis³⁹ and has been reflected in several prediction systems.⁴⁰ Methods designed to account for the specific characteristics of particular domain types should be considered an integral part of prediction algorithms (Fig. 2e).

According to previous studies, that amino acid variations that are caused by single nucleotide polymorphisms ("possible" variations) are slightly less deleterious than variations that occur when two or three nucleotides within the affected triplet are substituted ("impossible" variations).¹⁵ Although the likelihood of impossible variations occurring was low, we identified 97 (1.5%) of these variations within the analyzed dataset. Among impossible variations, we did not observe a significant improvement in the resolution of DAVs and NPAVs (Kruskal-Wallis one-way ANOVA on ranks, with Dunn's post-tests, $p > 0.05$ each). The DAVs were equally frequent among impossible (71%) and possible (68%) variations (χ^2 test $p > 0.05$ when the data were normalized to the total number of impossible variations) (Fig. 2g).

Because the effects of DAVs were not predicted by arbitrary thresholds, but by gene-specific thresholds (Figs. 1 and 3), we hypothesized that the prediction methods would differentiate between multiple diseases caused by variations in a single protein. Dunn's and Tukey's post-tests indicated the possibility of such differential diagnoses in several proteins (see Table S5 for an overview of outputs of statistical tests). We plotted the EVmutation and SNAP2 prediction scores for DAVs in nine proteins, for which the variations associated with the multiple phenotypes statistically differ (Fig. 3a-3i), and for two proteins (*GCK* and *HNF4A*) in which variations cause opposite phenotypes, i.e., diabetes and hyperglycemia (Fig. 3j-3k) or erythrocytosis and anemia (Fig. 3l). Despite the statistically significant differences, the variability in predictions of the genes prevented the

assignment of the variations to particular diseases, except for extreme values. Examples are listed below: a) The EVmutation score of *DMD* >-7 predicts muscular dystrophy of the Becker type (Fig. 3a). b) Noonan syndrome with multiple lentiginos is associated with variations with an EVmutation score for *PTPN11* <-4 and a SNAP2 score for *PTPN11* >30 (Fig. 3e). c) The EV mutation score for *UROD* >-4 or the SNAP2 score for *UROD* <0 predict the manifestation of porphyria cutanea tarda instead of hepatoerythropoietic porphyria (Fig. 3i).

Identification of variations under negative selection

We then used the newly obtained evidence-based knowledge to predict theoretically possible variations that have never been encountered in the clinic. This approach might highlight critical constrained variations that have not yet been linked to human disease phenotypes. Some of these variations likely exhibit such extreme constraints because they lead to extreme developmental disorders, are embryonically lethal or cause a long-term selection pressure by decreasing the fitness of their carriers. Although the theoretical ratio of impossible to possible variations was 2.47:1, the clinically observed ratio was 0.0143:1. The impossible and possible variations differed significantly in the scores obtained from both predictors (*t*-test $p < 0.001$ each), with EVmutation scores reaching -6.00 ± 2.42 and -4.83 ± 2.49 , and SNAP2 scores reaching 40 ± 51 and 18 ± 56 for impossible and possible variations, respectively. The gene-specific comparisons of the distribution of scores of impossible and possible variations and their comparison with the distribution of clinically documented DAVs and NPAVs are provided in Fig. S5.

The previous single-gene-oriented case study identified the potential existence of a pool of underrepresented variations in both healthy and disease-affected variation carriers.⁸ Since the present study provides the first large-scale adjustment of prediction scores based on clinical data, we focused on the detection of variations undergoing negative selection during molecular evolution. When performing this analysis (and in contrast to the aforementioned case study)⁸, we excluded any variations considered impossible by Bromberg *et al.*¹⁵ and analyzed the similarities of distributions of DAVs and possible theoretical variations. For simplicity, we compared the positions of the 10th percentiles for EVmutation scores and 90th percentiles for SNAP2 scores, which represent the predictions of amino acid changes with the most deleterious effects on proteins. Since possible theoretical variations include both putative DAVs and NPAVs, we expected that the analyzed values

calculated based on possible theoretical variations should be closer towards the scores of NPAVs. The differences in the 10th percentiles of EVmutation scores ranged from -1.093 to 3.360 (mean 0.921) and the differences in the 90th percentiles of SNAP2 scores ranged from -25.0 to 2.6 (mean -11.5).

In three genes (*PTPN11*, *HBB* and *G6PD*), the positions of 10th percentiles of the EVmutation scores were lower for DAVs than possible theoretical variations in the same genes. Similarly, in three genes (again *G6PD*, but also *HNF4A* and *EDA*) the positions of 90th percentiles of the SNAP2 scores were higher for DAVs than possible theoretical variations in the same genes. Thus, the variations that were predicted to be the most deleterious by EVmutation and/or SNAP2 were substantially depleted among DAVs compared to the spectra of possible theoretical variations in the same genes. These variations were therefore underrepresented among disease-affected variation carriers (Fig. 4a-f) and were under negative selection during molecular evolution. The heatmap of analyzed proteins, which were sorted according to the likelihood that their variations included variations under negative selection during molecular evolution, is shown in Fig. 4g. The phenotypes that are commonly associated with variations in these five genes are listed in Table 2. Confirmation of the negative selection against the underrepresented variations should consist of a series of studies that would compare the *in vitro* or *in vivo* effects of theoretical variations, which were hypothesized to be under negative selection, with clinically observed variations, which were within the range that did not seem to be subject to negative selection. During the peer-review of this manuscript, Havrilla *et al.*⁴¹ published a detailed map of constrained coding regions (CCR) in human genes and revealed that the most constrained regions are located in known disease loci. The genes encoding proteins associated with Mendelian diseases that we identified by applying the 10th/90th percentiles of DAVs partially overlapped with genes that ranked highly in the study by Havrilla *et al.*⁴¹ Namely, the CCR percentiles were 95.2% – 97.8% for *PTPN11* and 97.8% for *HNF4A*. However, other genes, namely *HBB*, *G6PD* and *EDA* were not among top hits in the previous CCR study.

Validation and conclusions

We validated the threshold values for EVmutation scores that were suggested in the proposed model. We established an independent dataset of variations in genes associated with Mendelian diseases (Tables S8-S9). The tested variations were classified according to ClinVar. The mean EVmutation scores for pathogenic and

benign variations were consistently below their previously suggested zero threshold (Table S10). The shift of the general EVmutation threshold to -2.13 led to a similar and significant improvement in the specificity of predictions of benign and likely benign variations, while the sensitivity remained higher than 96% for the pathogenic variations (Fig. 5a).

We calculated the sensitivity and specificity of the predictions retrieved from REVEL to determine whether the issue of low specificity was specifically associated with the outcomes of individual computational algorithms, such as EVmutation, or whether it also affected the data obtained from state-of-the-art consensus classifiers.⁵ REVEL predictions exhibited similar issues to the individual predictors. The scores for DAVs and NPAVs were gene-specific (Fig. 5b). The specificity was both low and gene-specific (Fig. 5c). Thus, although despite the consensus classifiers have the potential to eliminate the errors generated by individual predictors, they were prone to the systemic issue of low specificity.

All studies of human variations have a limitation in terms of how the variations are classified. For example, the incomplete penetrance may cause errors in the classification of rare variations.⁴² We re-analyzed the EVmutation and SNAP2 scores based on the ACMG criteria for the classification of variations to corroborate the key outcomes of the present study (Fig. 5d).²⁴ Variations classified as pathogenic according to the ACMG criteria were identified in both the DAV and NPAV datasets. EVmutation and SNAP2 identified only the first of these two groups as pathogenic. This difference in predictions was absent for common and rare variations among the NPAVs, which may reflect possible bias in the training or testing datasets for both of these methods.^{14,20}

The outcomes of prediction methods are often uncritically used, particularly by non-specialists in the field, who benefit from their use for the purpose of narrowing the number of hits identified during omics screens performed for scientific or clinical purposes. The uncritical use of the prediction methods is facilitated by including them in the tools commonly used for these purposes, such as the inclusion of SIFT and PolyPhen algorithms in the Ensembl genome browser (<http://www.ensembl.org/>; Release 90 cited). Based on accumulating evidence, the prediction methods are often over-interpreted, mainly because they exhibit high false positive rates,^{8,43} and sufficiently complex datasets used for the design, testing and training of the

methods are lacking.⁴⁴ Any distinct effects observed at the molecular level depend on the context and can be compensated by intrinsic regulatory pathways of the organism, which particularly applies to the effects of variations in nonessential peripheral enzymes and signaling proteins.^{14,45-46}

New prediction methods are rapidly released, and EVmutation is one of the most recent contributions to the field.¹⁴ EVmutation is important because it includes epistasis when modeling the effect of the respective variation. We provided the first match for the EVmutation (and SNAP2 and PoPMuSiC 2.1) prediction outcomes with clinical phenotypes of a large pool of pathogenic and benign variations in genes associated with Mendelian diseases. EVmutation, similar to the other tested prediction methods, had high sensitivity but also extremely low specificity. We suggested the use of evidence-based thresholds, which were obtained by calculating and testing several variants of the thresholds until we reached 98.6% sensitivity and 83.6% specificity, leaving the certain pool of variations unresolved (if needed, the size of this pool can be decreased at the cost of decreasing the sensitivity and/or specificity). The predictions provided better resolution for variations located in enzymes and predominantly those within enzymatic domains. For some proteins, the use of numerical outputs of predictions combined with evidence-based thresholds distinguished between multiple diseases caused by variations in the same protein. We identified large previously unreported pools of variations that underwent negative selection during molecular evolution and were absent in patients. These variations were particularly prominent in *G6PD*, *PTPN11*, *HNFA4A* and *HBB*. Further research should focus on the use of evidence-based thresholds for categories of variations defined using the Human Phenotype Ontology (such as the Phenomizer or Phevor)⁴⁷⁻⁴⁸ and phenome-wide association studies (PheWAS).⁴⁹⁻⁵⁰

Based on the large-scale analysis provided in the present study, we suggest the use of evidence-based thresholds to improve the outcomes of any prediction methods that produce numerical scores. Improved settings of the individual methods will facilitate the outcomes of consensus classifiers represented by REVEL⁵, PredictSNP⁵¹, PredictSNP2⁵², CADD⁵³ or DANN⁵⁴. The evolutionary variation analysis approach described here is the first to enable the highly specific identification of likely disease-causing missense variations that have not yet been associated with any clinical phenotype.

Acknowledgements: We thank Petr Šimčík (www.petrsimcik.cz) for his help with data mining. The study was supported by the Czech Science Foundation project 15-03834Y and Charles University in Prague projects Primus/MED/32, GA UK 1428218 and 260387/SVV/2017. The authors declare that they have no conflicts of interest. All financial support for the study was acknowledged.

Competing interests: DS and PH have been funded by the Czech Science Foundation project 15-03834Y and Charles University in Prague projects Primus/MED/32, GA UK 1428218 and 260387/SVV/2017. The authors declare no other competing interests.

Data and materials availability: All data are available in the main text or in the supplementary materials.

Funding: DS and PH have been funded by the Czech Science Foundation project 15-03834Y and Charles University in Prague projects Primus/MED/32, GA UK 1428218 and 260387/SVV/2017. The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author contributions: PH and DS conceived and designed the experiments, acquired and analyzed the data, wrote the paper, are responsible for the integrity of this work, revised the article's intellectual content and approved the final version.

REFERENCES

- 1 Biesecker, L. G. & Green, R. C. Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.* **371**, 1170 (2014).
- 2 Simm, F., *et al.* Identification of SLC20A1 and SLC15A4 among other genes as potential risk factors for combined pituitary hormone deficiency. *Genet. Med.* **20**, 728–736 (2018).
- 3 Tennessen, J. A., *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
- 4 The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- 5 Ioannidis, N. M., *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
- 6 Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
- 7 Bamshad, M. J., *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
- 8 Šimčíková, D., Kocková, L., Vackářová, K., Těšínský, M. & Heneberg, P. Evidence-based tailoring of bioinformatics approaches to optimize methods that predict the effects of nonsynonymous amino acid substitutions in glucokinase. *Sci. Rep.* **7**, 9499 (2017).
- 9 Hayat, S., Sander, C., Marks, D. S. & Elofsson, A. All-atom 3D structure prediction of transmembrane β -barrel proteins from sequences. *Proc. Nat. Acad. Sci. U. S. A.* **110**, 5413–5418 (2015).
- 10 Wang, Y. & Barth, P. Evolutionary-guided de novo structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy. *Nat. Commun.* **6**, 7196 (2015).
- 11 Peled, S., *et al.* De-novo protein function prediction using DNA binding and RNA binding proteins as a test case. *Nat. Commun.* **7**, 13424 (2016).
- 12 Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).

- 13 Dehouck, Y., Kwasigroch, J. M., Gilis, D. & Rooman, M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinform.* **12**, 151 (2011).
- 14 Hopf, T. A., *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
- 15 Bromberg, Y., Kahn, P. C. & Rost, B. Neutral and weakly nonneutral sequence variants may define individuality. *Proc. Nat. Acad. Sci. U. S. A.* **110**, 14255–14260 (2013).
- 16 Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
- 17 Libbrecht, M. W. Machine learning in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
- 18 Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucl. Acid. Res.* **43**, W7–W14 (2015).
- 19 Adebali, O., Reznik, A. O., Ory, D. S. & Zhulin, I. B. Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations. *Genet. Med.* **18**, 1029–1036 (2016).
- 20 Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genom.* **16**(Suppl 8), S1 (2015).
- 21 DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6**, 678–687 (2005).
- 22 de Visser, J. A. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).
- 23 Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).
- 24 Nykamp, K., *et al.* Sherlock: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet. Med.* **19**, 1105–1117 (2017).
- 25 Mathe, E., *et al.* Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucl. Acid. Res.* **34**, 1317–1325 (2006).
- 26 Richards, S., *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

- 27 Stenson, P. D., *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).
- 28 Liu, L., *et al.* High-density SNP genotyping to define beta-globin locus haplotypes. *Blood Cells Mol. Dis.* **42**, 16–24 (2009).
- 29 Steele, A. M., *et al.* The previously reported T342P *GCK* missense variant is not a pathogenic mutation causing MODY. *Diabetologia* **54**, 2202–2205 (2011).
- 30 Chellapa, K., *et al.* Src tyrosine kinase phosphorylation of nuclear receptor HNF4 α correlates with isoform-specific loss of HNF4 α in human colon cancer. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2302–2307 (2012).
- 31 Houlleberghs, H., *et al.* Oligonucleotide-directed mutagenesis screen to identify pathogenic Lynch syndrome-associated *MSH2* DNA mismatch repair gene variants. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4128–4133 (2016).
- 32 Maxwell, K. N., *et al.* Evaluation of ACMG-guideline based variant classification of cancer susceptibility and non-cancer-associated genes in families affected by breast cancer. *Am. J. Hum. Genet.* **98**, 801–817 (2016).
- 33 Walsh, R., *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* **19**, 192–203 (2016).
- 34 Hicks, S., Wheeler, D. A., Plon, S. E. & Kimmel, M. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum. Mutat.* **32**, 661–668 (2011).
- 35 Riera, C., Padilla, N. & de la Cruz, X. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. *Hum. Mutat.* **37**, 1013–1024 (2016).
- 36 Pawson, T. Protein modules and signaling networks. *Nature* **373**, 573–580 (1995).
- 37 Aronson, H. E., Royer, W. E. & Hendrickson, W. A. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Protein Sci.* **3**, 1706–1711 (1994).
- 38 Rost, B. Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318**, 595–608 (2002).
- 39 Miller, M. L., *et al.* Pan-cancer analysis of mutation hotspots in protein domains. *Cell Systems* **1**, 197–209 (2015).

- 40 Salgado, D., *et al.* UMD-Predictor: a high-throughput sequencing compliant system for pathogenicity prediction of any human cDNA substitution. *Hum. Mutat.* **37**, 439–446 (2016).
- 41 Havrilla, J. M., Pedersen, B. S., Layer, R. M., Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
- 42 Bastarache, L., *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233–1239(2018).
- 43 Romeo, S., *et al.* Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J. Clin. Invest.* **119**, 70–79 (2009).
- 44 Rost, B., Radivojac, P. & Bromberg, Y. Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett.* **590**, 2327–2341 (2016).
- 45 Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
- 46 Boucher, J. I., Bolon, D. N. & Tawfik, D. S. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Sci.* **25**, 1219–1226 (2016).
- 47 Singleton, M. V., *et al.* Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.* **94**, 599–610 (2014).
- 48 Bone, W. P., *et al.* Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet. Med.* **18**, 608–617 (2016).
- 49 Simonti, C. N., *et al.* The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**, 737–741 (2016).
- 50 Posey, J. E., *et al.* Resolution of disease phenotypes resulting from multilocus genomic variation. *N. Engl. J. Med.* **376**, 21–31 (2017).
- 51 Bendl, J., *et al.* PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* **10**, e1003440 (2014).
- 52 Bendl, J., *et al.* PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. *PLoS Computat. Biol.* **12**, e1004962 (2016).

- 53 Kircher, M., *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- 54 Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinform.* **31**, 761–763 (2015).
- 55 Sarkozy, A., *et al.* Clinical and molecular analysis of 30 patients with multiple lentigines LEOPARD syndrome. *J. Med. Genet.* **41**, e68 (2004).
- 56 Yoshida, R., *et al.* Two novel and one recurrent *PTPN11* mutations in LEOPARD syndrome. *Am. J. Med. Genet. A* **130A**, 432–434 (2004).
- 57 Osawa, R., *et al.* A novel *PTPN11* missense mutation in a patient with LEOPARD syndrome. *Br. J. Dermatol.* **161**, 1202–1204 (2009).
- 58 Digilio, M. C., *et al.* Grouping of multiple-lentigines/LEOPARD and Noonan syndromes on the *PTPN11* gene. *Am. J. Hum. Genet.* **71**, 389–394 (2002).
- 59 Seishima, M., *et al.* Malignant melanoma in a woman with LEOPARD syndrome: identification of a germline *PTPN11* mutation and a somatic *BRAF* mutation. *Br. J. Dermatol.* **157**, 1297–1299 (2007).
- 60 Conti, E., *et al.* A novel *PTPN11* mutation in LEOPARD syndrome. *Hum. Mutat.* **21**, 654 (2003).
- 61 Keren, B., *et al.* *PTPN11* mutations in patients with LEOPARD syndrome: a French multicentric experience. *J. Med. Genet.* **41**, e117 (2004).
- 62 Sarkozy, A., *et al.* Correlation between *PTPN11* gene mutations and congenital heart defects in Noonan and LEOPARD syndromes. *J. Med. Genet.* **40**, 704–708 (2003).
- 63 Atik, T., *et al.* Mutation spectrum and phenotypic features in Noonan syndrome with *PTPN11* mutations: definition of two novel mutations. *Indian J. Pediatr.* **83**, 517–521 (2016).
- 64 Tartaglia, M., *et al.* *PTPN11* mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. *Am. J. Hum. Genet.* **70**, 1555–1563 (2002).
- 65 Al-Gazali, L. & Ali, B. R. Mutations of a country: a mutation review of single gene disorders in the United Arab Emirates (UAE). *Hum. Mutat.* **31**, 505–520 (2010).
- 66 Knott, M., *et al.* Novel and Mediterranean beta thalassemia mutations in the indigenous Northern Ireland population. *Blood Cells Mol. Dis.* **36**, 265–268 (2006).

- 67 Colah, R., *et al.* Regional heterogeneity of beta-thalassemia mutations in the multi ethnic Indian population. *Blood Cells Mol. Dis.* **42**, 241–246 (2009).
- 68 Villegas, A., *et al.* Hb Santander [beta34(B16)Val --> Asp (GTC --> GAC)]: a new unstable variant found as a de novo mutation in a Spanish patient. *Hemoglobin* **27**, 31–35 (2003).
- 69 Henderson, S. J., *et al.* Ten years of routine α - and β -globin gene sequencing in UK hemoglobinopathy referrals reveals 60 novel mutations. *Hemoglobin* **40**, 75–84 (2016).
- 70 Zanella-Cleon, I., *et al.* Strategy for identification by mass spectrometry of a new human hemoglobin variant with two mutations in Cis in the beta-globin chain: Hb S-Clichy [beta6(A3)Glu-->Val; beta8(A5)Lys-->Thr]. *Hemoglobin* **33**, 177–187 (2009).
- 71 Wajcman, H., *et al.* Two new hemoglobin variants with increased oxygen affinity: Hb Nantes [beta34(B16)Val-->Leu] and Hb Vexin [beta116(G18)His-->Leu]. *Hemoglobin* **27**, 191–199 (2003).
- 72 McClure, R. F., Hoyer, J. D. & Mai, M. The JAK2 V617F mutation is absent in patients with erythrocytosis due to high oxygen affinity hemoglobin variants. *Hemoglobin* **30**, 487–489 (2006).
- 73 Shin, S. Y., Bang, S. M. & Kim, H. J. A novel hemoglobin variant associated with congenital erythrocytosis: Hb Seoul [β 86(F2)Ala→Thr] (HBB:c.259G>A). *Ann. Clin. Lab. Sci.* **46**, 312–314 (2016).
- 74 Vulliamy, T., Beutler, E. & Luzzatto, L. Variants of glucose-6-phosphate dehydrogenase are due to missense mutations spread throughout the coding region of the gene. *Hum. Mutat.* **2**, 159–167 (1993).
- 75 Bulliamy, T., Luzzatto, L., Hirono, A. & Beutler, E. Hematologically important mutations: glucose-6-phosphate dehydrogenase. *Blood Cells Mol. Dis.* **23**, 302–313 (1997).
- 76 Yan, T., *et al.* Incidence and complete molecular characterization of glucose-6-phosphate dehydrogenase deficiency in the Guangxi Zhuang autonomous region of southern China: description of four novel mutations. *Haematologica* **91**, 1321–1328 (2006).
- 77 McGlacken-Byrne, S. M., *et al.* The evolving course of *HNF4A* hyperinsulinaemic hypoglycaemia--a case series. *Diabet. Med.* **31**, e1–e5 (2014).
- 78 Flanagan, S.E., *et al.* Diazoxide-responsive hyperinsulinemic hypoglycemia caused by *HNF4A* gene mutations. *Eur. J. Endocrinol.* **162**, 987–992 (2010).

- 79 Colclough, K., Bellanne-Chantelot, C., Saint-Martin, C., Flanagan, S. E. & Ellard, S. Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha and 4 alpha in maturity-onset diabetes of the young and hyperinsulinemic hypoglycemia. *Hum. Mutat.* **34**, 669–685 (2013).
- 80 Urbanová, J., *et al.* Positivity for islet cell autoantibodies in patients with monogenic diabetes is associated with later diabetes onset and higher HbA1c level. *Diabet. Med.* **31**, 466–471 (2014).
- 81 Harries, L.W., *et al.* The diabetic phenotype in HNF4A mutation carriers is moderated by the expression of HNF4A isoforms from the P1 promoter during fetal development. *Diabetes* **57**, 1745–1752 (2008).
- 82 Song, S., *et al.* EDA gene mutations underlie non-syndromic oligodontia. *J. Dent. Res.* **88**, 126–131 (2009).
- 83 Lee, K. E., *et al.* Oligodontia and curly hair occur with ectodysplasin-a mutations. *J. Dent. Res.* **93**, 371–375 (2014).
- 84 Ruiz-Heiland, G., *et al.* Novel missense mutation in the EDA gene in a family affected by oligodontia. *J. Orofac. Orthop.* **77**, 31–38 (2016).
- 85 Cluzeau, C., *et al.* Only four genes (*EDA1*, *EDAR*, *EDARADD*, and *WNT10A*) account for 90% of hypohidrotic/anhidrotic ectodermal dysplasia cases. *Hum. Mutat.* **32**, 70–72 (2011).
- 86 Guazzarotti, L., *et al.* Phenotypic heterogeneity and mutational spectrum in a cohort of 45 Italian males subjects with X-linked ectodermal dysplasia. *Clin. Genet.* **87**, 338–342 (2015).
- 87 Clauss, F., *et al.* X-linked and autosomal recessive Hypohidrotic Ectodermal Dysplasia: genotypic-dental phenotypic findings. *Clin. Genet.* **78**, 257–266 (2010).
- 88 Monreal, A.W., Zonana, J. & Ferguson, B. Identification of a new splice form of the *EDA1* gene permits detection of nearly all X-linked hypohidrotic ectodermal dysplasia mutations. *Am. J. Hum. Genet.* **63**, 380–389 (1998).
- 89 Schneider, P., *et al.* Mutations leading to X-linked hypohidrotic ectodermal dysplasia affect three major functional domains in the tumor necrosis factor family member ectodysplasin-A. *J. Biol. Chem.* **276**, 18819–18827 (2001).
- 90 Pääkkönen, K., *et al.* The mutation spectrum of the *EDA* gene in X-linked anhidrotic ectodermal dysplasia. *Hum. Mutat.* **17**, 349 (2001).

Figure legends

Fig. 1. The efficiency of the EVmutation prediction method in predicting the effects of missense variations with known clinical phenotype on proteins known to cause for Mendelian diseases. (a) Flowchart showing the sources and approaches used for data retrieval, the construction of datasets and subsequent analyses. The selection of analyzed genes associated with Mendelian diseases was based on combined information retrieved from the Human Gene Mutation Database (HGMD), UniProtKB/Swiss-Prot, Protein Data Bank (PDB) and Online Mendelian Inheritance in Man (OMIM). Information about disease associations and no-phenotype associations of clinically observed variations was retrieved from the ClinVar database and the Ensembl browser. Additional information about proteins (domains) and variations (frequency) was obtained from the Pfam database and the Exome Aggregation Consortium (ExAC) browser, respectively. A vertical line indicates the arbitrary threshold for variations with an effect. (b) The distribution of numerical EVmutation scores calculated for missense variations with known clinical phenotypes. (c) The relative percentage of correct predictions of disease and no clinical phenotypes using EVmutation scores calculated for the 44 analyzed proteins. (d) The distribution of numerical EVmutation scores calculated for disease-associated and no phenotype-associated missense variations with known clinical phenotypes in 44 proteins that cause Mendelian diseases sorted according to the evolutionary conservation of affected amino acids in mammals. Conserved amino acids ($GV = 0$) were conserved in all ten examined mammalian orthologs. Variable amino acids ($GV > 0$) were not conserved in at least one of the ten examined mammalian orthologs of the respective protein.

Fig. 2. The predictions differ for evolutionarily conserved proteins, such as AR or PTEN, for variations within and outside of protein domains and for enzymes and proteins without enzymatic functions. (a) Evolutionary divergence of the amino acid sequences of AR and PTEN reported as the number of amino acid substitutions per site by averaging all sequence pairs between primates and other groups. (b-c) GV scores for amino acids within the AR (b) and PTEN (c) sequences. The data are shown separately for GV scores calculated based on mammalian protein orthologs (the two lines at the zero GV score) and extended MSAs that included more evolutionarily distant taxa. The data are shown for disease-associated and no phenotype-associated variations. Relative ranks among tested variations are shown to reflect the different numbers of variations included in each analyzed group. (d) EVmutation and SNAP2 scores applied to disease-associated and no phenotype-

associated variations that are present or absent from protein domains. Data are presented as medians \pm SD. (e) Differences in median EVmutation and SNAP2 scores between disease-associated and no phenotype-associated variations located within the indicated protein domains. Abbreviations for the domains: AGAL, alpha-galactosidase A; ATCase/OTCase, aspartate/ornithine carbamoyltransferase, carbamoyl-P binding and Asp/Orn binding domains; CPOX, coproporphyrinogen III oxidase; DHE1, dehydrogenase E1 component; FRNADBD, ferric reductase, NAD binding domain; GTPCH, GTP cyclohydrolase I; G6PDH, glucose-6-phosphate dehydrogenase, NAD binding and C-terminal domains; HXK, hexokinase_1 and hexokinase_2; LBDNHR, ligand-binding domain of nuclear hormone receptor; PK, protein kinase; PTK, protein tyrosine kinase; PTP SH2, Src Homology 2 domain. (f) Median EVmutation and SNAP2 scores calculated for disease-associated and no phenotype-associated variations in the four indicated enzyme classes and in proteins without enzymatic functions. (g) EVmutation and SNAP2 scores calculated for disease-associated and no phenotype-associated variations considered possible or impossible variations according to Bromberg *et al.*¹⁵ Data are shown as medians \pm SD.

Fig. 3. The efficiency of the prediction methods in discriminating among multiple diseases caused by missense variations in the indicated proteins. EVmutation and SNAP2 scores are shown for proteins with significantly different disease-specific scores (a-i) or that result in the opposite phenotypes (j-l). (a) *DMD*, (b) *ELANE*, (c) *FLNA*, (d) *HPRT1*, (e) *PTPN11*, (f) *RET*, (g) *TGFRB2*, (h) *TP63*, (i) *UROD*, (j) *GCK*, (k) *HNF4A*, and (l) *HBB*.

Fig. 4. The detection of variations under negative selection during molecular evolution: an example of the application of evidence-based knowledge. (a-f) The distribution of observed disease-associated variations compared to the distribution of possible¹⁰ theoretical variations. The data are shown for the proteins for which negative values were obtained from the calculation of the differences in the 10th percentiles of EVmutation scores – (a) *PTPN11*, (b) *HBB* and (c) *G6PD* – and for genes for which positive values were obtained from the calculation of the differences in 90th percentiles of SNAP2 scores – (d) *G6PD*, (e) *HNF4A* and (f) *EDA*. (g) The heatmap of proteins causing Mendelian diseases sorted according to the likelihood that their variations included variations that were under negative selection during molecular evolution. Ranges of differences in median values: -1.093 – 3.36 (EVmutation) and -25 – 2.6 (SNAP2).

Fig. 5. Validation of the model, identification of the specificity of the consensus classifier REVEL, and the application of the American College of Medical Genetics and Genomics (ACMG) criteria for the classification of variations. (a) Validation of the threshold values for EVmutation that were suggested in the proposed model. Validation was performed using a set of 1723 variations in 63 genes (Tables S8-S10), which were classified according to ClinVar. The data are presented as relative percentages of correct predictions using the arbitrary EVmutation threshold (0.00), the evidence-based threshold that allows 95% sensitivity (-2.13) and the threshold that allows 95% specificity (-8.81). (b-c) REVEL, a consensus classifier, is associated with the issue of low specificity, similar to the individual computational algorithms. REVEL scores were retrieved for a set of 2721 variations in 21 genes. Mean REVEL scores for the individual genes discriminated well between the disease-associated and no phenotype-associated variations (b). However, because a large overlap in the predictions was observed, the specificity was low for most of the analyzed genes (c). Data are presented (b) as the means \pm SE or (c) as relative percentages of correct predictions of the association of the variations with diseases (upper row) or no phenotypes (lower row). (d) Application of the ACMG criteria for the classification of variations, which classify the variations as benign (1B and higher) and pathogenic (0.5 P and higher) according to the population frequencies of the variations (Table S11). The EVmutation and SNAP2 scores were analyzed separately for the disease- and no phenotype-associated variations. Data are shown as means \pm SE.

List of Tables

Table 1. The key used to assign of the clinically observed variations. Abbreviations used: DIS – disease-associated; PART – partial phenotype-associated; NO PHEN – no phenotype-associated; EXCL – excluded ambiguous data.

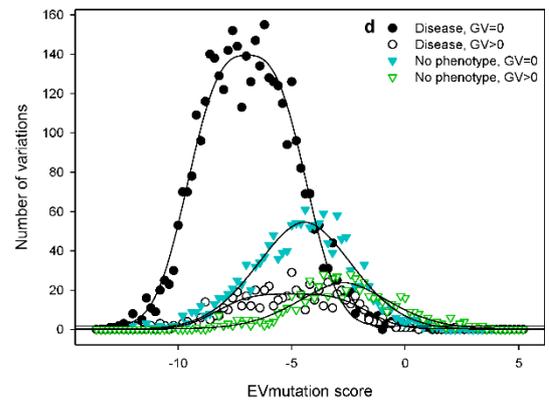
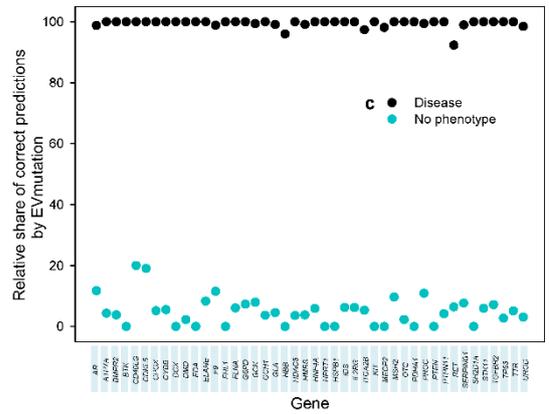
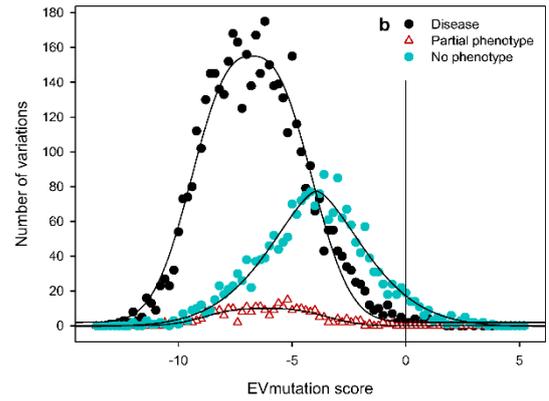
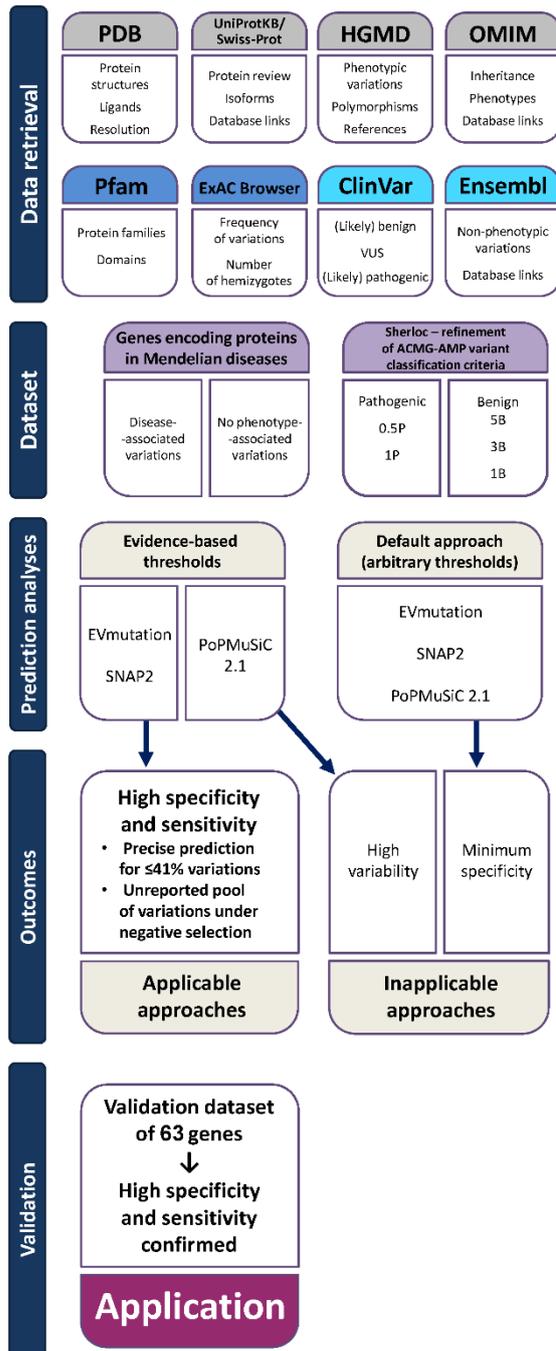
1a)	In HGMD, the variation is absent.	2
1b)	In HGMD, the variation is present, but causes “no phenotype” according to dbSNP.	NO PHEN
1c)	In HGMD, the variation is present and is defined as a “disease causing mutation”.	4
1d)	In HGMD, the variation is present but has with definitions other than those listed in 1b) and 1c)	2
2a)	In ClinVar, the variation is present and defined as “benign”, “likely benign” or “variants of uncertain significance” (VUSs).	NO PHEN
2b) 2a).	In ClinVar, the variation is absent or present, with definitions other than those listed in 2a).	3
3a)	In Ensembl, the variation is present but has no associated phenotype.	NO PHEN
3b)	In Ensembl, the variation is present and associated with a phenotype.	5
4a)	In ClinVar, the variation is present and defined as “benign” or “likely benign”.	EXCL
4b)	In ClinVar, the variation is present but not defined as “benign” or “likely benign”.	5
5a)	In HGMD, all variations classified as “disease-causing mutations” within the respective gene are associated with a single disease or syndrome with a Mendelian inheritance pattern.	DIS
5b)	In HGMD, the variations classified as “disease-causing mutations” within the respective gene are associated with two diseases with a Mendelian inheritance pattern, one caused by the activating and the other by inactivating variations (e.g., erythrocytosis vs anemia).	DIS

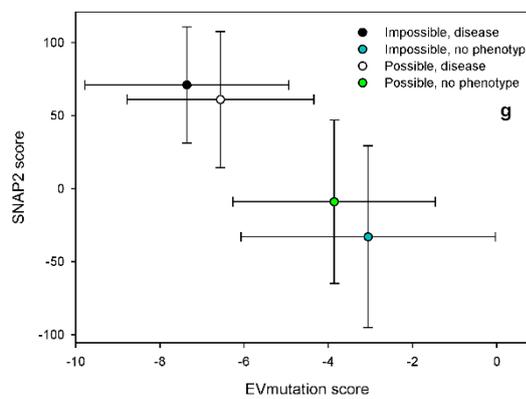
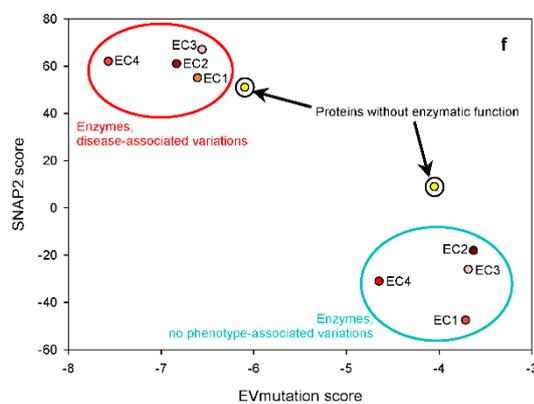
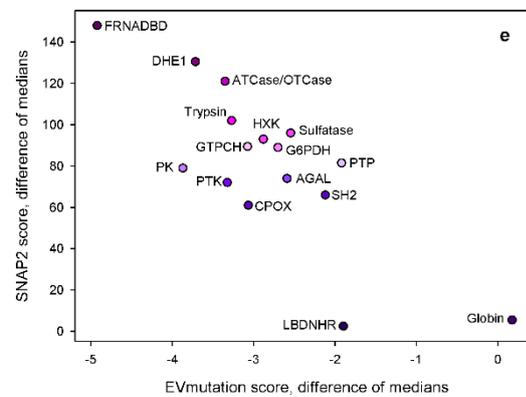
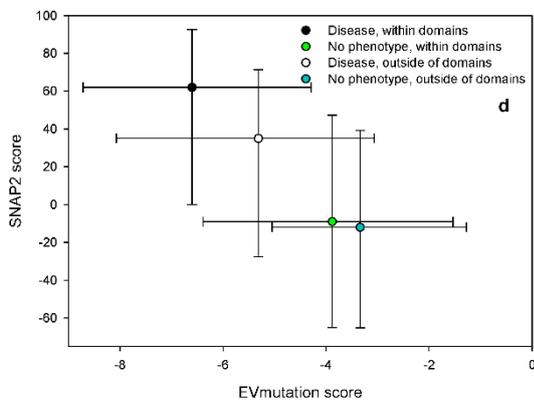
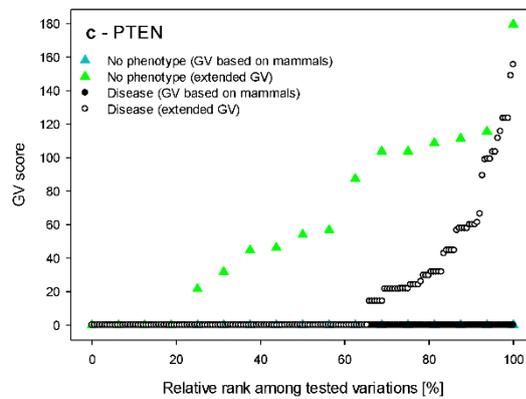
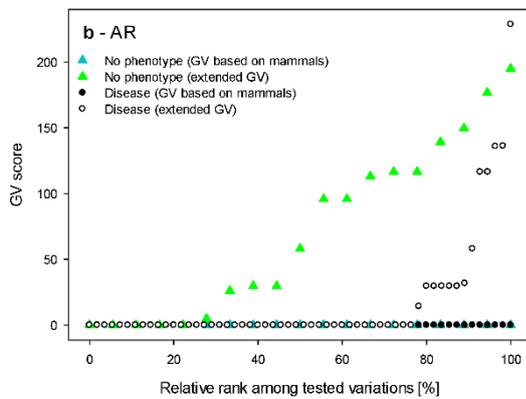
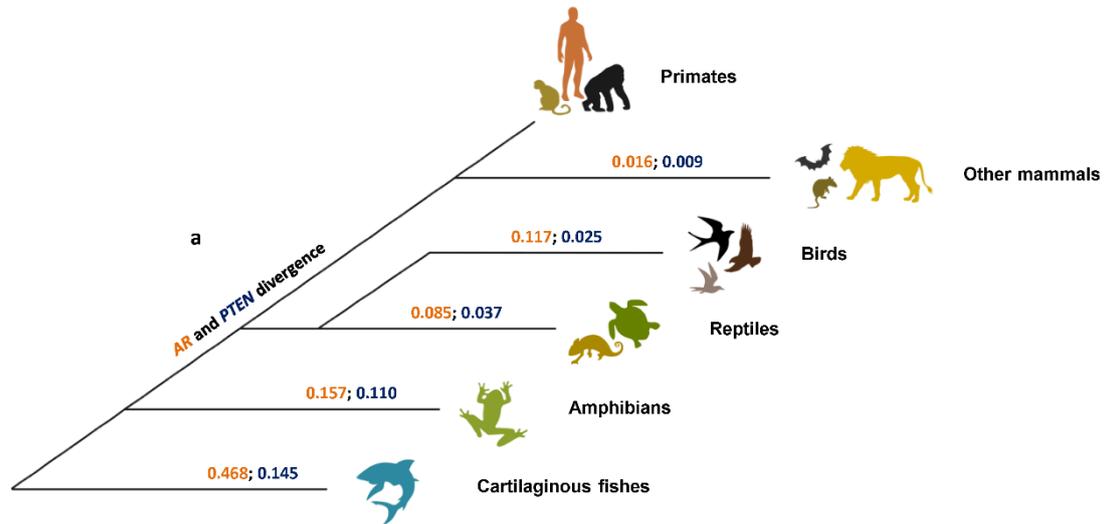
-
- 5c) In HGMD, the variations classified as “disease-causing mutations” within the respective gene are associated with two diseases with a Mendelian inheritance pattern, both of which are caused by variations exerting similar effects with a different intensity (e.g., Menkes syndrome vs occipital horn syndrome or Duchenne vs Becker muscular dystrophy); variations cause a complete phenotype. DIS
- 5d) In HGMD, the variations classified as “disease-causing mutations” within the respective gene are associated with two diseases with a Mendelian inheritance pattern, both of which are caused by variations exerting similar effects with a different intensity (e.g., Menkes syndrome vs occipital horn syndrome or Duchenne vs Becker muscular dystrophy); variations cause the less pathological phenotype. PART
-

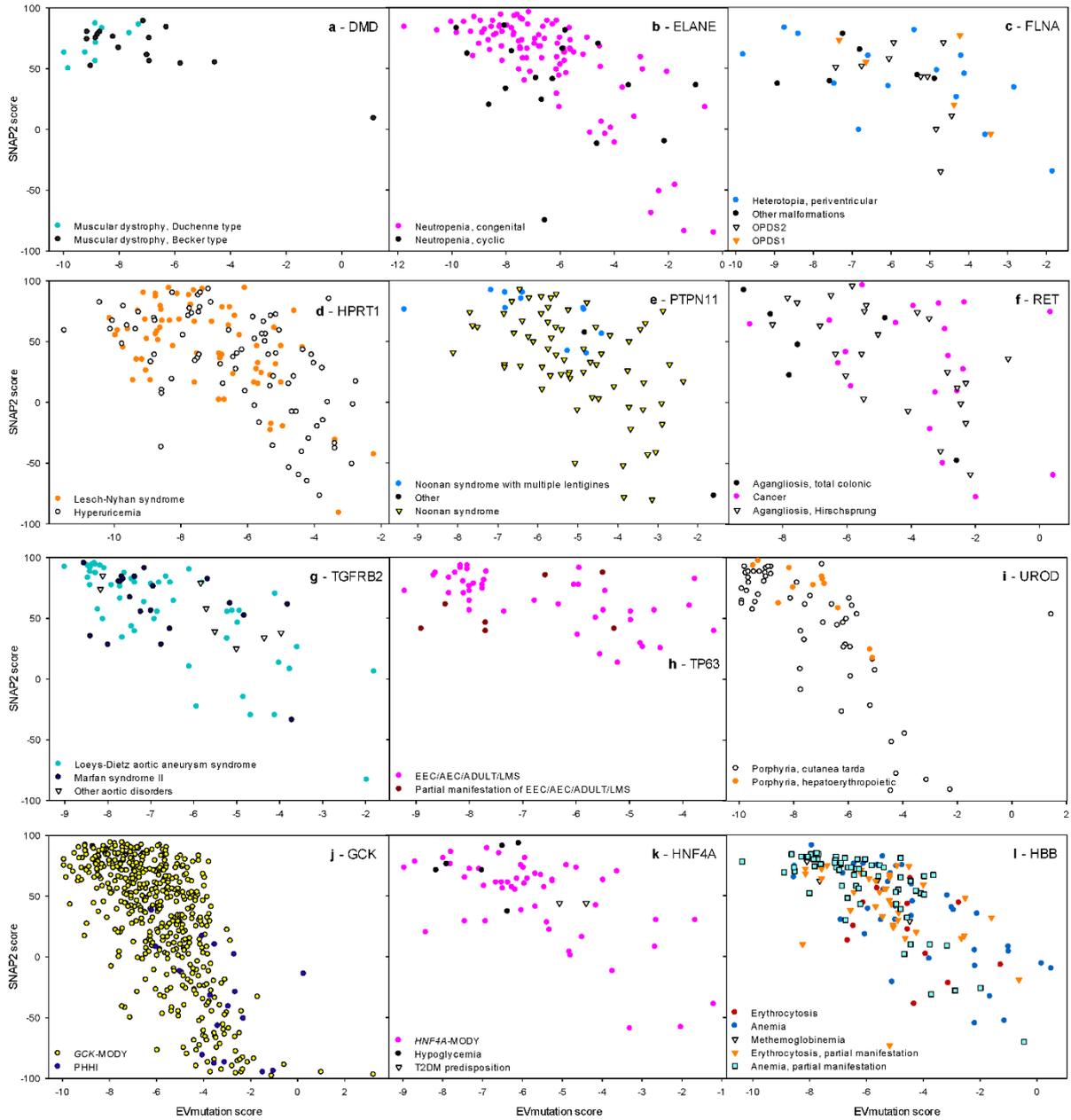
Table 2. Major phenotypes associated with genes that were underrepresented among disease-affected carriers. See Table S7 for a complete list of phenotypes associated with analyzed variations and source references.

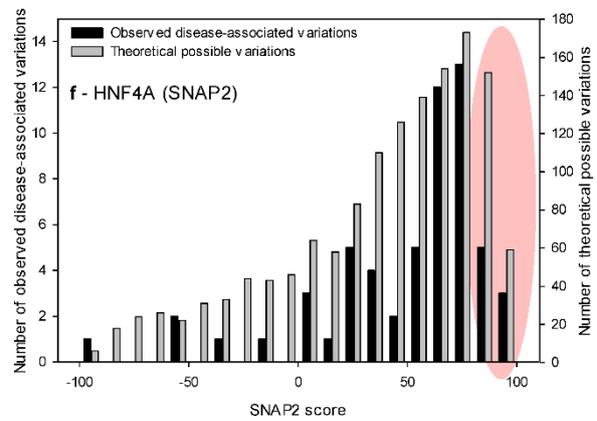
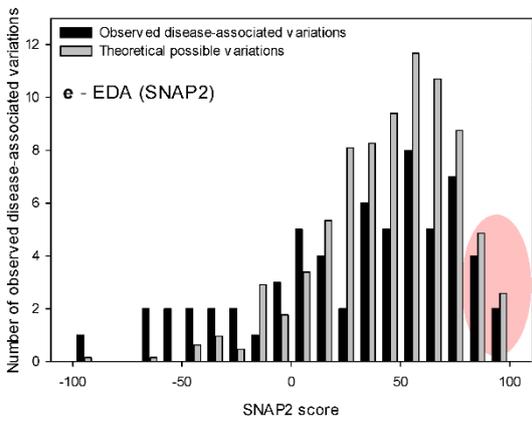
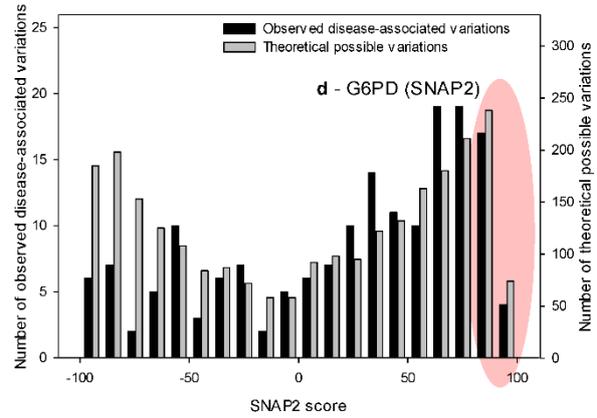
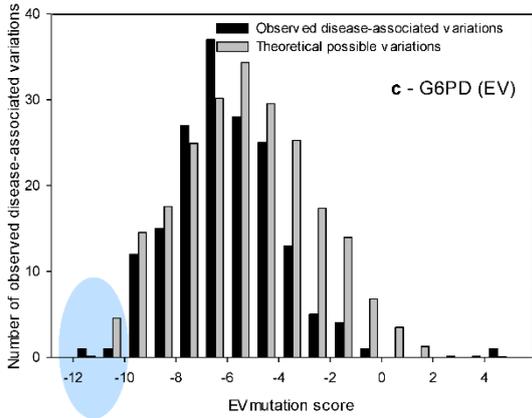
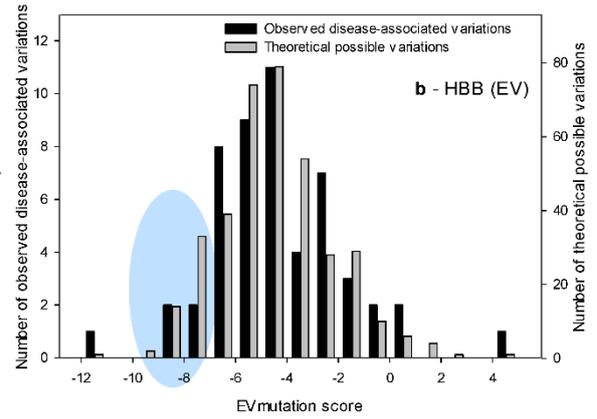
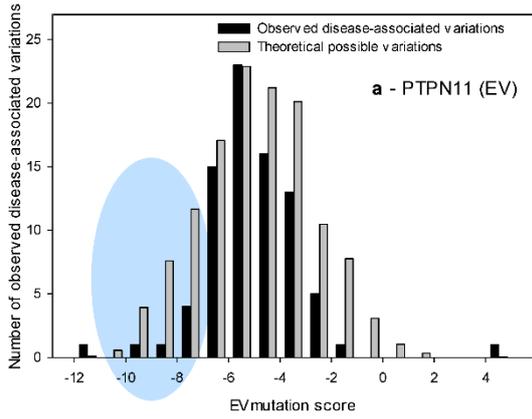
Gene	Phenotype	References
<i>PTPN11</i>	Multiple lentigines / LEOPARD syndrome	55-61
	Noonan syndrome	62-64
<i>HBB</i>	Thalassaemia beta	65-67
	Hemolytic anemia	68-70
	Erythrocytosis	71-73
<i>G6PD</i>	Glucose-6-phosphate dehydrogenase deficiency	74-76
<i>HNF4A</i>	Hypoglycemia, hyperinsulinemic	77-79
	Diabetes, <i>HNF4A</i> -MODY	79-81
<i>EDA</i>	Oligodontia	82-84
	Ectodermal dysplasia, hypohidrotic	85-87
	Ectodermal dysplasia	88-90

a

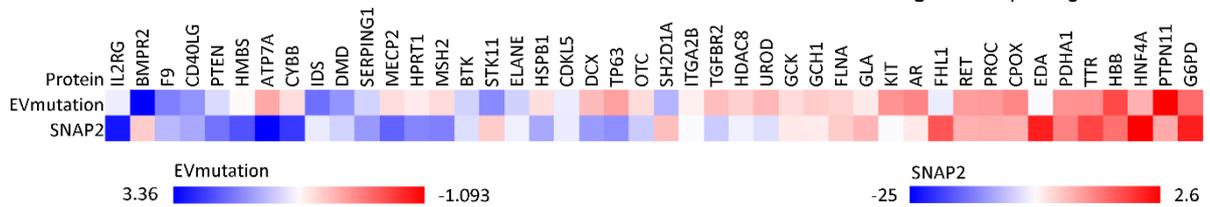


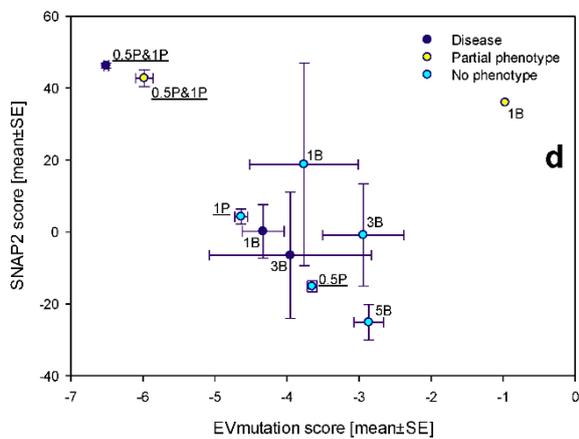
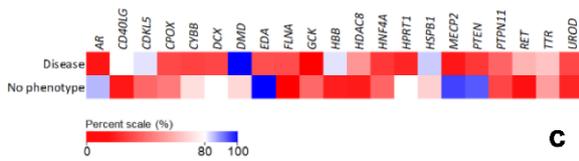
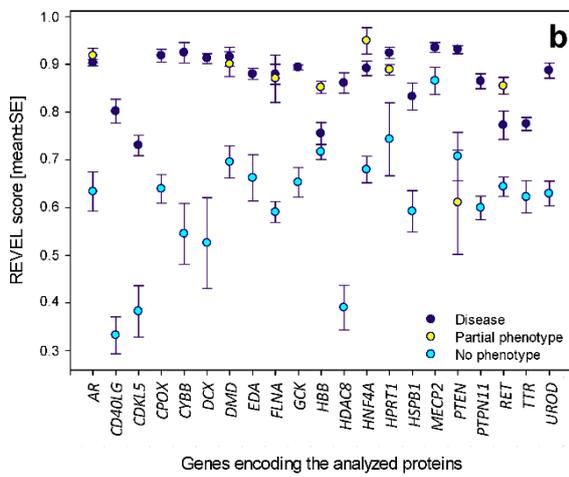
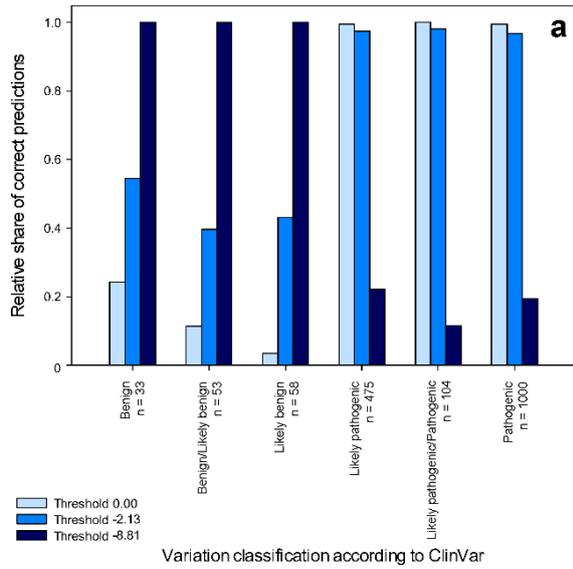






g - Heatmap of negative selection





Daniela Šimčíková, Dominik Gardáš, Tomáš Pelikán, Petr Heneberg

Isoform-specific roles of HK1 in ovarian cancer cells

Unpublished (2019)

Restriction analysis of CRISPR/Cas9 clones

Using the PCR combined with restriction analysis, we revealed that the chosen CRISPR/Cas9 approach allowed the generation of several putative *HK1* KO clones of the HEK293T and TOV-112D cells (Fig. 2). We further verified the CRISPR/Cas9-induced changes using bidirectional Sanger sequencing, which confirmed that the targeted sites were cleaved by Cas9 and inaccurately repaired, thereby confirming HK knockouts (Fig. 3). We used the *HK1* KO clones verified by the restriction analysis and Sanger sequencing for further experiments.

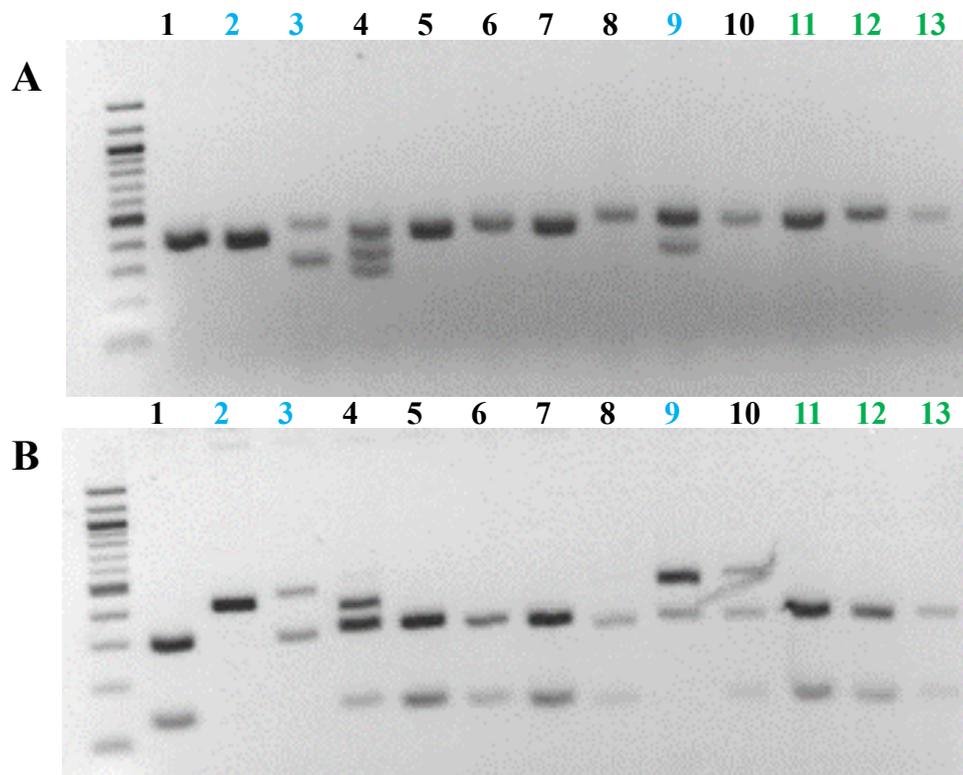


Fig. 2. PCR of CRISPR/Cas9 HK1 KO HEK293T clones performed with primers for the sgRNA22 target site (A) and restriction cleavage of these PCR products by EcoRI (B). The HEK293T clones highlighted in green were transfected with the empty plasmid pSpCas9(BB)-2A-GFP (without a sgRNA; serving as controls for CRISPR/Cas9 KO experiment). The HEK293T clones highlighted in blue displayed differences after restriction cleavage, thus they were sequenced and subjected to Western blotting with the anti-HK1 antibody.

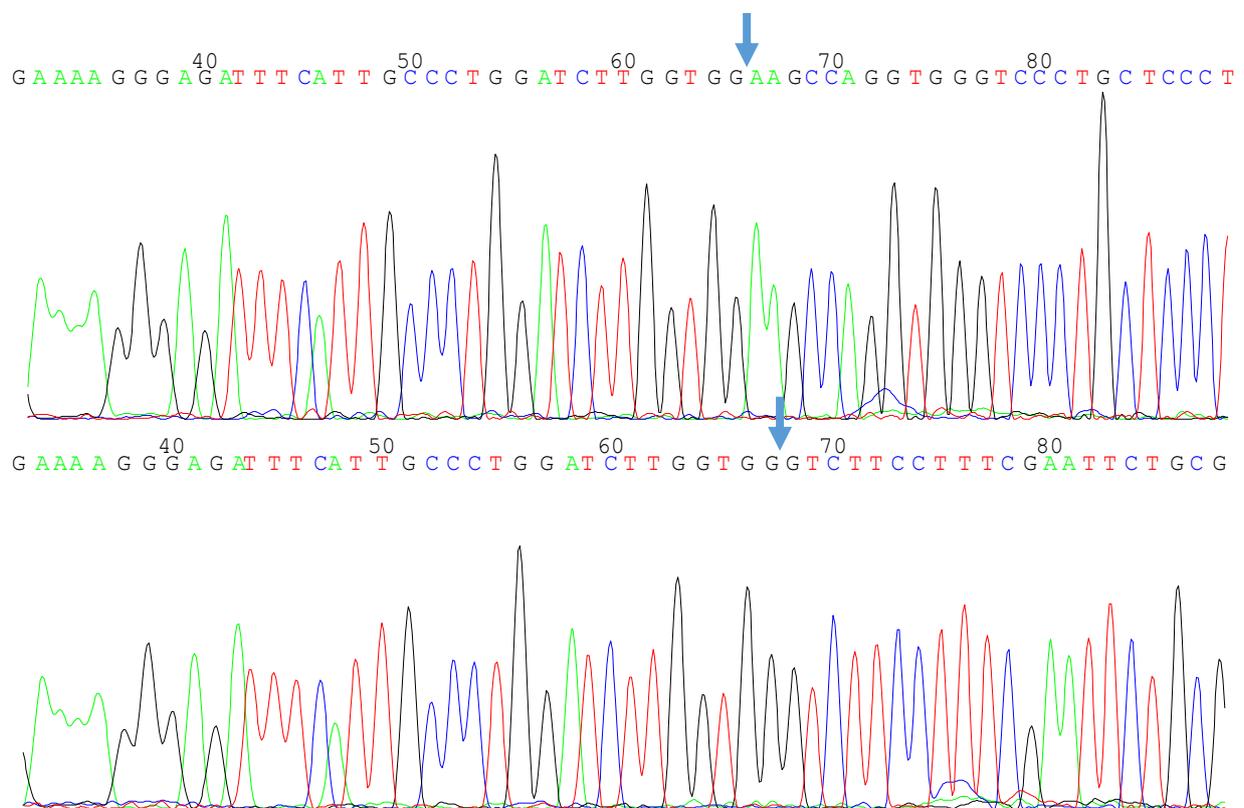


Fig. 3. Representative electropherograms of Sanger sequencing. The DNA sequence affected by sgRNA16-guided Cas9 resulting in a large deletion (upper sequence) and the wild-type DNA sequence (lower sequence). The blue arrows show the beginning of deletion.

Western blot analysis

We confirmed the CRISPR/Cas9-induced HK1 or HK2 deletions in newly generated cell clones by Western blotting using HK1- and HK2-specific antibodies, respectively (Fig. 4). The results of the Western blotting analysis corresponded to those that were obtained by using the restriction cleavage.

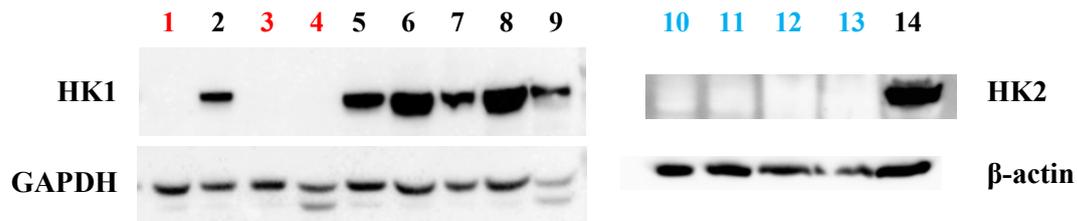


Fig. 4. Western blot analysis of TOV-112D CRISPR/Cas9 HK KO clones were tested for HK1 (1 - 9) or HK2 (10 - 14) expression. The confirmed HK1 KO clones are under red numbers (1, 3, 4) and the confirmed HK2 KO clones are under blue numbers (10 – 13).

Metabolic enzymes mapping

In the HK1 KO clone E9-14-3, we observed twofold elevated expression of LDHA and MTCO2 (complex IV of ETC), which has been confirmed by both rabbit antibody and mouse antibody in the WB antibody cocktail for ETC (Fig. 5). Changes of expression levels of other glycolytic enzymes (HK2, PFKP, PGAM-1 and PKM2) as well as other complexes of ETC (I, II, III and V) were negligible (Fig. 5). We used two commonly recommended loading controls, vinculin and β -actin. However, vinculin appeared to be downregulated (from two- to fourfold lower expression) in the HK1 KO clone, thus we rather calculated intensity of the bands according to the bands of β -actin.

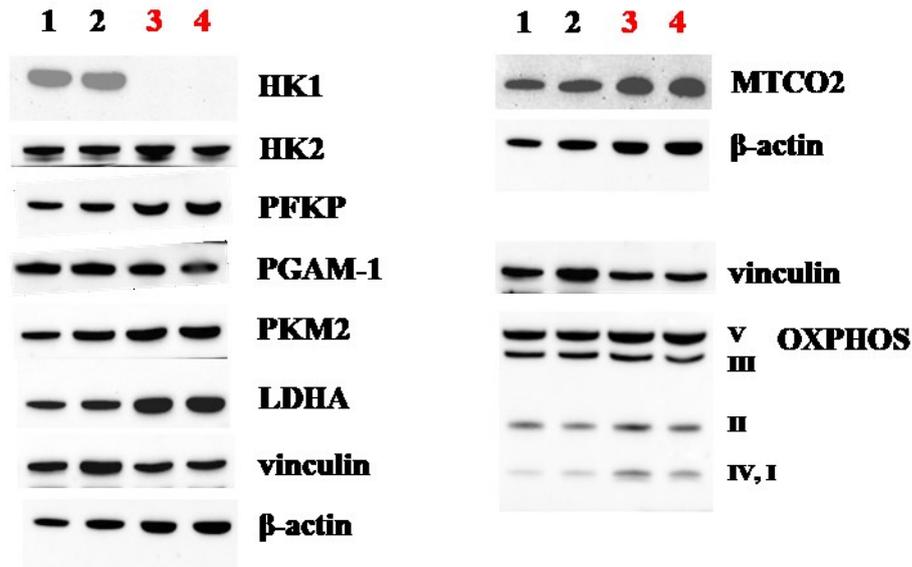


Fig. 5. Western blotting of proteins involved in glycolysis and electron transport chain in the CRISPR/Cas9 clones D11-14-5 (control clone - 1, 2) and E9-14-3 (HK1 KO clone - 3, 4). The cells were cultivated in DMEM with 1 g/L glucose (2, 4) and 4.5 g/L glucose (1, 3).

Signaling pathways mapping

In the HK1 KO clone E9-14-3, we observed approximately fivefold increased phosphorylation of Rictor at Thr1135 by p70 S6K, which negatively regulates mTORC2 as a part of negative feedback mechanism controlling PKB/Akt activity (*Julien et al., 2010; Treins et al., 2010*). This finding has been supported by ninefold increased phosphorylation of PKB/Akt at Ser473 (Fig. 6), which is known to be caused by mTORC2 (*Sarbassov et al., 2005*). Concurrently, we did not observed changes in AMPK activation in HK1 KO cells, although Raptor was phosphorylated by AMPK, thereby inhibiting mTORC1 (Fig. 6).

On the other hand, mTORC1 appeared to be still active, since p70 S6K is remarkably phosphorylated at Thr389 (approximately fivefold) which is considered a hallmark of mTORC1 activation (*Burnett et al., 1998*). This finding is confirmed by the above-mentioned phosphorylation of Rictor in mTORC2. Moreover, the increased phosphorylation of 4E-BP1 at

Thr37, Thr70 and slightly at Ser65 suggested activation of mTORC1. These observations together with threefold elevated phosphorylation of S6RP likely leads to increased translation in the cells (Fig. 6). Concurrently, we observed remarkably increased expression of the oncogene *c-Myc* in HK1 KO cells, thereby promoting cell growth and proliferation in accordance with activated PKB/Akt and mTORC1 (Fig. 7).

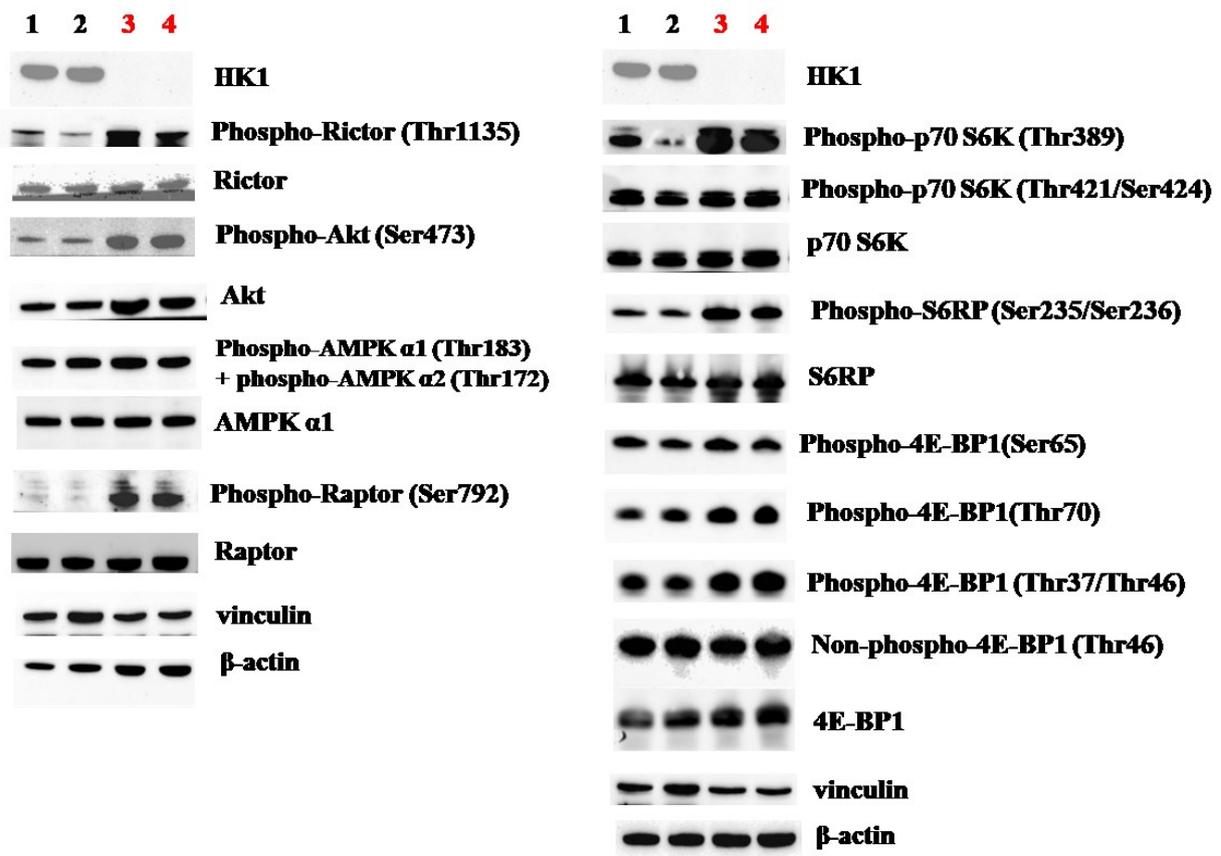


Fig. 6. Western blotting of proteins upstream from and included in mTORC1 (on the left), and proteins downstream from mTORC1 (on the right) chain in the CRISPR/Cas9 clones D11-14-5 (control clone - 1, 2) and E9-14-3 (HK1 KO clone - 3, 4). The cells were cultivated in DMEM with 1 g/L glucose (2, 4) and 4.5 g/L glucose (1, 3).

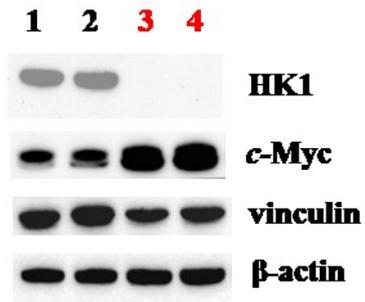


Fig. 7. Western blotting of *c-Myc* chain in the CRISPR/Cas9 clones D11-14-5 (control clone - 1, 2) and E9-14-3 (HK1 KO clone - 3, 4). The cells were cultivated in DMEM with 1 g/L glucose (2, 4) and 4.5 g/L glucose (1, 3).

DISCUSSION

Precise and personalized medicine enables tailoring of treatment for individual patients with taking into account their genetic backgrounds. Availability of comprehensive genome databases and lowering costs of genetic examinations lead us to the question how to interpret this data in the context of the whole organism. How could we tackle with functional analyses for all these genomic data? For this purpose, researchers have been developing prediction algorithms, in which they have used experimental data paired with clinical, biophysical and/or evolutionary information in order to extrapolate patterns from their outcomes and apply them on newly found variations. In our work, we tried to imply outcomes of prediction algorithms in some Mendelian diseases.

First, we have focused on monogenic diabetes caused by inactivating variations within the GCK molecule, also known as *GCK-MODY*. Particularly, we have chosen variations, which were previously found in Czech patients suffering from hyperglycemia and subsequently being diagnosed as MODY patients. For this purpose, we implemented an enzyme assay and purification procedure for the recombinant GCK expressed from *E. coli*, which is used and accepted among research groups that study the effects of *GCK-MODY* variations (*Davis et al., 1999*). In the accordance with the previously published data (*Gloyn et al., 2005; Sagen et al., 2006*), some of these variations, namely R250C and C434Y, displayed normal kinetics, although their association with MODY have been reported from several independent Czech families with confirmed family history. The amino acid exchange at Cys434 affects one of the experimentally confirmed nitrosylation sites within the GCK molecule (*Rizzo & Piston, 2003*). Although the function of nitrosylation at Cys434 is unknown (as opposed to the nitrosylation of Cys371), the variation C434Y found in four independent Czech families with MODY phenotype appears to cause MODY (*Pruhova et al., 2010*). Furthermore, R250C is associated with a more severe phenotype that manifests during childhood and was confirmed in MODY

patients of Serbian and Czech origin (*Pinterova et al., 2007; Milenkovic et al., 2008*). On the other hand, R250C has been predicted as deleterious by prediction algorithms, which we employed for extended analysis of their use in personalized medicine.

The prediction algorithms are to be an indispensable part of research with potential in diagnostics, especially in time of generating large datasets quite easily due to next-generation sequencing and other omics studies, for which complete functional analyses are mostly unfeasible and ineffective. Even widely used databases use outcomes of some prediction algorithms, such as the SIFT and PolyPhen algorithms providing predictions listed in the Ensembl genome browser (www.ensembl.org). These two algorithms have been frequently used in studies for investigation of particular proteins, including GCK. Some of these studies presented experimental data in agreement with the two algorithms (*Steele et al., 2011*), although another study has brought evidence of their high false positive rates (29% for SIFT and 43% for PolyPhen) and low rates of correct predictions (53% for SIFT and 63% for PolyPhen) (*Romeo et al., 2009*). Therefore, their outcomes may be rather interpreted taking into account experimental data, whereas the conclusions that would be based solely on predictions could be misleading. We have extended our study with epistatic approach incorporated in the EVmutation method (*Hopf et al., 2017*). The authors of EVmutation claimed that this method outperforms the commonly used SIFT and PolyPhen. Nevertheless, we found out that EVmutation is also associated with poor sensitivity for activating and neutral variations in GCK, and its sensitivity for inactivating variations did not excel over SIFT and PolyPhen significantly.

We have realized that the use of evidence-based thresholds may overcome low specificity in order to distinguish up to 75% of *GCK*-*MODY*-associated variations from *GCK* variations associated with hypoglycemia and normoglycemia. However, these activating and neutral *GCK* variations could not be identified selectively by any of the prediction methods,

since their outcomes for these variation groups have largely been overlapped. Not surprisingly, studies on variations in MODY-associated genes have revealed similar problems, such as very low specificity of SIFT and PolyPhen. The most authoritative study investigated activating and deactivating variations in GCK, ABCC8 and KCNJ11 (*Flanagan et al., 2010*). They found that sensitivity of SIFT and PolyPhen reached 69% and 68%, respectively, whereas specificity was only 13% and 16%, respectively (*Flanagan et al., 2010*). Other two studies have shown false predictions of SIFT and PolyPhen on GKRP variations (*Johansen et al., 2010; Rees et al., 2012*).

The results obtained using a single protein (GCK) stimulated us to check, whether the same issues are associated with other proteins that are associated with Mendelian diseases. For the follow-up study, we have assembled and curated two non-overlapping large databases of clinical phenotypes caused by missense variations in 44 and 63 genes associated with Mendelian diseases. We used these databases to establish and validate the model allowing to improve the predictions of clinical phenotypes caused by missense variations by the prediction algorithms with numerical (therefore scalable) outcomes. It was important to exclude the algorithms that generate binary responses (such as SIFT or PolyPhen), since we would not be able to tailor their predictions, unless being able to change their code. In contrast, the algorithms with numerical outcomes allow simple changes of the threshold according to the available evidence for the pathogenicity of variations. To verify the reliability of this analysis, we tested outcomes by one of the state-of-the-art consensus classifiers, REVEL, which turned to be subject to similar issues as the individual computational approaches. In summary, we proposed the evidence-based approach that allows modifying the settings of prediction methods in a way that they generate predictions of clinical phenotypes with both high sensitivity and specificity. This adjustment cannot be done with the predictors that are integrated into the Ensembl genome browser (SIFT and PolyPhen). However, SIFT and PolyPhen do not outperform the analysed

computational approaches even under ad hoc settings, and, of course, they are not superior to them under the modified settings (*Simcikova et al., 2017*).

The newly proposed prediction approach (*Simcikova & Heneberg, subm.*) is being far from optimal, but, so far, it is the first approach that allows providing specific predictions without loss of sensitivity. We confirmed that even a simple shift of the threshold in the approaches, such as EVmutation and SNAP2, is associated with improved predictions. These adjustable thresholds should be applied when using these methods and should be incorporated in consensus classifiers, since it may increase their reliability. Further, we point to the fact that the thresholds differ for different classes of proteins, which was not reflected so far in any generalized suggestions for the use of predictors. However, we could not have predicted all the variations correctly, since many variations belong to the “grey zone”, which is hardly predictable.

To move the topic forward towards tumorigenesis and cancer metabolism, our further efforts focused on a group of somatic cancer-associated variations in GCK. We have retrieved these variations from the COSMIC database. We found that a subset of somatic cancer-associated variations were activating and/or increasing protein stability, similarly as in the case of variations causing PHHI. Regarding tertiary structure of GCK, these activating variations are concentrated in or near the heterotropic allosteric activator site (*Gloyn et al., 2003*). This site is distinct from the substrate-binding cleft for glucose and ATP considered to be potential drug targeting site for the treatment of type 2 diabetes (*Gloyn et al., 2003*). In contrast, neutral or inhibitory variations are distributed randomly across the GCK molecule. Instead of cooperative binding of glucose according to the Hill kinetics, all activating variations displayed rather Michaelis-Menten kinetics and/or decreased perception to the competitive inhibitor *N*-acetylglucosamine (GlcNAc).

The clustering of activating somatic cancer-associated variations in GCK resembled a focal distribution of known activating cancer-associated variations in proto-oncogenes, such as *TP53* (Kato *et al.*, 2003) or *BRAF* (Cantwell-Dorris *et al.*, 2011). These activating variations were present in the region, which consists of amino acids 151 – 180 and is essential for regulation of GCK cooperativity (Gloyn *et al.*, 2003). It has been shown that variations in this region may suppress completely GCK cooperativity, thus promoting rapid GCK activation (Whittington *et al.*, 2015). Other of activating variations were located in the heterotropic allosteric activator site and its surroundings as mentioned-above. Furthermore, the activating variations also decreased GCK cooperativity towards the Michaelis-Menten kinetics.

Although some of somatic GCK variations associated with cancer displayed increasing activity and stability, which could be advantageous for tumor growth, we did not observe any supportive evidence for this hypothesis and began to pay our attention on other hexokinases involved in tumorigenesis. First, we retrieved and analysed transcriptomics data from Expression Atlas (<https://www.ebi.ac.uk/gxa/home>). We compared amounts of *HK1* and *HK2* transcripts in cancer cell lines and found out that some cancer cell lines prefer *HK1*, instead of *HK2*, although *HK2* has been generally considered a preferential isoform in tumors (Nakashima *et al.*, 1986).

To analyse the roles of *HK1* and *HK2*, we selected ovarian cancer cell lines, in which the ratio of *HK1*/*HK2* transcripts was strongly skewed towards *HK1*, unlike normal ovarian tissue with *HK2* mRNA expression is higher than that of *HK1* (Fig. 8). Subsequently, we prepared *HK1* and *HK2* knockout ovarian cancer cells by CRISPR/Cas9 and investigated changes in metabolic and signaling pathways.

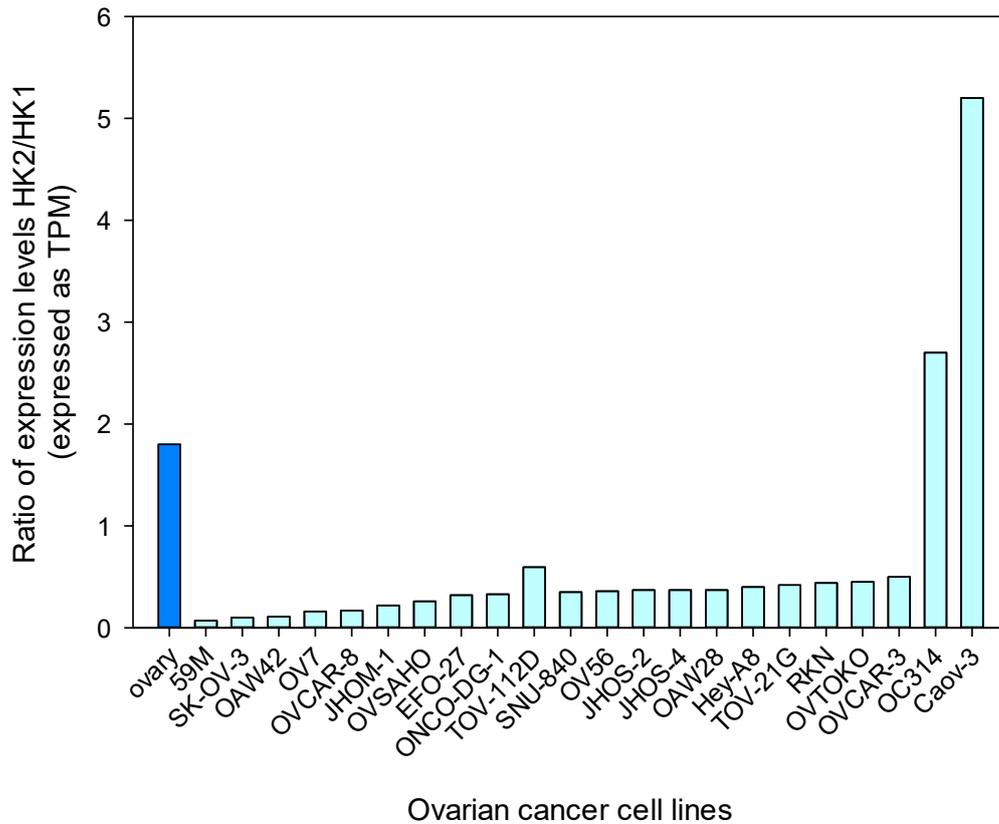


Fig. 8. Ratio of mRNA levels of *HK2/HK1* expressed by TPM (transcripts per million) in normal ovary tissue and cancer cell lines derived from ovary tissue. Data was retrieved from Expression Atlas. Data on *HK1* and *HK2* indicated high transcription levels compared with *HK3* and *GCK*.

Concerning glycolytic enzymes, we observed significantly increased expression of LDHA. In contrast, expression of *HK2*, *PFKP*, *PGAM-1* and *PKM2* remained unaffected. Previously, the LDHA upregulation was reported to be associated with *c-Myc* transactivation (Shim *et al.*, 1997; Dang *et al.*, 2008). *c-Myc* may promote the Warburg effect by upregulation of LDHA, which leads to the lactate overproduction and increased clonogenicity, e.g., in the model of Burkitt's lymphoma cells (Shim *et al.*, 1997). Therefore, we checked *c-Myc* in the

examined clones and, indeed, we confirmed the elevated *c-Myc* expression in the HK1 KO cells.

Besides LDHA, *c-Myc* has been shown to upregulate HK2, PFKP and PKM2 (*Kim et al., 2007; Yap et al., 2011; Gupta et al., 2018*). However, at the model of ovarian cancer cell lines, we did not observe any changes in the expression of PFKP and PKM2, and we even did not observe any changes in the expression levels of HK2. These findings were unexpected given that HK2 serves as an isoenzyme of HK1, which shares the identical enzymatic function. The finding that HK2 expression is not increased in response to loss of HK1 in the examined ovarian cancer cell lines suggests that the glucose phosphorylation does not have the gate-keeping function in the maintenance of growth rate in the nutrient-rich medium. In agreement with the above, we observed similar growth rates of HK1 knockout and control cells. The decline of glucose concentration in the medium from 4.5 g/L to 1 g/L glucose did not alter these conclusions.

Importantly, only a subset of *c-Myc*-target genes is induced in any experimental system or condition. The responses of target genes following *c-Myc* activation are likely to depend on a variety of other factors and change with cell type and environment (*Fernandez et al., 2003*). Available evidence suggests that there may exist a feedback loop between HK1 and *c-Myc*, as *c-Myc* binds on the promoter region of *HK1* gene *in vitro* (*Ciribilli et al., 2016*). However, the regulation of HK1 expression is not straightforward, since only oscillating levels of HK1 corresponding to glucose consumption can be observed in the cells with inactivated *c-Myc*, whereas the presence of active *c-Myc* does not alter the HK1 expression in these cells (*Altman et al., 2015*).

In addition to changes in *c-Myc*, we observed the elevated levels of proteins involved in ETC, particularly cytochrome *c* oxidase (complex IV), in the HK1 knockout cells. This finding may be also explained by upregulation of *c-Myc*, since *c-Myc* promotes mitochondrial

biogenesis and upregulates the gene *CYCS* encoding cytochrome *c*, which transfers electrons to the complex IV (*Li et al., 2005*). In our further study, we aim to investigate other metabolic enzymes and pathways that are likely regulated by *c-Myc*, such as enolase 1 (ENO1) in glycolysis (*Osthus et al., 2000*), glutaminolysis (*Gao et al., 2009*) or lipid synthesis (*Morrish et al., 2010*). Furthermore, according to the first data, which we obtained from a pilot untargeted metabolomics experiment with HK1 knockout and control cells (data are not shown), we did not observe changes in lactate concentration. Therefore, we aim to perform more comprehensive metabolomics analysis that will focus on metabolites resulting from pathways affected by *c-Myc*.

We assumed that the HK1 deletion must have led to the disruption of nutrient balance, therefore, we investigated effects of HK1 deletion on the mechanistic target of rapamycin (mTOR) signaling pathway. mTOR coordinates cell growth and metabolism with environmental inputs, including nutrients and growth factors (*Saxton & Sabatini, 2017*). mTOR, a serine/threonine protein kinase, is a catalytic subunit of two distinct protein complexes, known as mTOR Complex 1 (mTORC1) and 2 (mTORC2). mTORC1 consists of three core components: mTOR, Raptor (regulatory protein associated with mTOR), and mLST8 (mammalian lethal with Sec13 protein 8, also known as G β L) (*Kim et al., 2002; Kim et al., 2003*). mTORC2, like mTORC1, contains mTOR and mLST8; however, instead of Raptor, mTORC2 contains Rictor (rapamycin insensitive companion of mTOR) (*Jacinto et al., 2004*). We have observed activation of mTORC2 in the HK1 KO cells, since PKB/Akt was phosphorylated significantly at the mTORC2 phosphorylation site Ser473 (*Sarbassov et al., 2005*). The phosphorylation sites at Ser473 by mTORC2 and Thr308 by PDK1 are required for maximal activation of PKB/Akt (*Alessi et al., 1996*).

The observed changes in the phosphorylation of mTORC complexes can be related to above-reported changes in *c-Myc* activity. In glioblastoma cells, it has been shown that

mTORC2 promotes inactivating phosphorylation of histone deacetylases, which results in acetylation of the transcription factors FoxO1 and FoxO3 and the subsequent release of *c-Myc* from a suppressive miR-34c-dependent network (Masui *et al.*, 2013). Apart from the PKB/Akt-independent mechanism of *c-Myc* regulation, mTORC2 promotes the release of *c-Myc* through phosphorylation of PKB/Akt (Peck *et al.*, 2013). The existence of the HK1 – *c-Myc* – mTORC1/2 axis requires further verification by alterations of expression or activity of its key members beyond HK1.

The signaling pathway upstream from mTORC1 involves AMPK, which has been considered a tumor suppressor due to inhibition of mTORC1. However, recent evidence suggests that AMPK may directly activate mTORC2, thereby promoting tumorigenesis (Kazyken *et al.*, 2019). We did not observe any HK1-induced changes of either AMPK phosphorylation or its upstream tumor suppressor LKB1 expression (data are not shown). Despite that, we observed phosphorylation of Raptor at Ser792 by AMPK, which should contribute to the inhibition of mTORC1, in the HK1 KO cells. However, we found that mTORC1 is still active in HK1 KO cells. mTORC1 downstream effectors, such as 4E-BP1 (Thr37, Ser65 and Thr70) or p70S6K (Thr389), were more strongly phosphorylated in HK1 KO cells, thereby promoting translation and anabolism (Saxton & Sabatini, 2017). Consistently, we observed elevated phosphorylation of Rictor at Thr1135, which is mediated by active p70S6K (Julien *et al.*, 2010; Treins *et al.*, 2010). Moreover, the significantly elevated phosphorylation of S6RP indicates higher translation activity in the HK1 KO cells (Ruvinsky *et al.*, 2005).

All these findings, except of phosphorylation of Raptor, suggest pro-survival and pro-growth effects in the HK1 KO cells. Notwithstanding, we aim to explore these dependencies in our further work on hexokinases. Apart from the changes in metabolism, the HK1 deletion was associated with downregulation of vinculin, which we initially aimed to use only as a loading

control. We hypothesize that loss of this adhesion-related protein promotes the epithelial-mesenchymal transition (EMT), a critical process in tumor invasion and metastasis (*Li et al., 2014*). Further, we aim to focus on other markers of EMT in order to verify these newly emerging roles of HK1 beyond glycolysis.

CONCLUSION

In the part that focused on GCK, we experimentally confirmed the causativity of GCK variations found in Czech patients with *GCK-MODY* by performing functional analysis of their GCK variations. Our results were consistent with the outcomes of the prediction algorithms, particularly SNAP2 and EV mutation. To improve specificity of prediction algorithms concerning other GCK variations, we suggested a model for tailoring numerical outcomes of the prediction algorithms. We further used and verified our model on comprehensive dataset of variations causing Mendelian diseases, thereby increasing specificity of prediction algorithms. Despite that, we found that large number of variations are still unpredictable even with our tailored approach. Further, we refined pH optimum of human GCK and HK2 and pointed out the undesired influence of ATP concentrations on buffering capacity of commonly used buffers.

In the part concerning tumorigenesis, we realized that a subset of somatic cancer-associated variations in GCK appeared to be advantageous for tumors, since these variations were activating and thermostable. On the other hand, we did not find more supportive evidence for a role of GCK in cancer. In contrast, we obtained results on the HK1 KO ovarian cancer cells that appeared to trigger pro-survival and pro-growth effects after loss of HK1. We observed increased expression of the oncogene *c-Myc* and LDHA as well as activation of mTOR complexes. Nevertheless, we must perform supportive experiments, which will enable us to make definitive conclusions.

LIST OF ABBREVIATIONS

ACMG	American College of Medical Genetics and Genomics
ADP	adenosine diphosphate
AMP	adenosine monophosphate
AMPK	AMP-activated protein kinase
AR	androgen receptor
ATP	adenosine triphosphate
bp	base pair
Cas	CRISPR-associated
CFTR	cystic fibrosis transmembrane conductor receptor
CRISPR	clustered regularly interspaced short palindromic repeats
CRISPRi	CRISPR interference
crRNA	CRISPR RNA
CTM4G	Charcot-Marie-Tooth disease type 4G
DAV	disease-associated variation
DMEM	Dulbecco's Modified Eagle Medium
DSB	double-strand break
DTT	dithiothreitol
EMT	epithelial-mesenchymal transition
ENO1	enolase 1
ETC	electron transport chain
ExAC	Exome Aggregation Consortium
FAH	fumarylacetate hydrolase
GCK	glucokinase

GD	Grantham deviation
GKRP	glucokinase regulatory protein
GlcNAc	<i>N</i> -acetylglucosamine
GO	Gene Ontology
GSIR-T	threshold for glucose-stimulated insulin release
GST	glutathione- <i>S</i> -transferase
GV	Grantham variation
HDR	homology-directed repair
HGMD	Human Gene Mutation Database
HIF	hypoxia-inducible factor
HK	hexokinase
HRP	horseradish peroxidase
IDH1	isocitrate dehydrogenase 1
IPTG	isopropyl β -D-1-thiogalactopyranoside
KO	knockout
LDHA	lactate dehydrogenase A
LKB1	liver kinase B1
miR	microRNA
MODY	maturity-onset diabetes of the young
MTCO2	mitochondrially encoded cytochrome <i>c</i> oxidase 2
mTOR	mechanistic target of rapamycin
mTORC	mTOR complex
NADP	nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
NHEJ	non-homologous end-joining

NPAV	no phenotype-associated variation
NSHA	non-spherocytic haemolytic anemia
OXPHOS	oxidative phosphorylation
PAM	protospacer adjacent motif
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PDAC	pancreatic ductal adenocarcinoma
PDB	Protein Data Bank
PDK1	phosphoinositide-dependent kinase-1
PEA15	phosphoprotein enriched in astrocytes
PFK	phosphofructokinase
PFKP	PFK, platelet
PGAM-1	phosphoglycerate mutase 1
PHHI	hyperinsulinemic hypoglycaemia of infancy
PI3K	phosphoinositide 3-kinase
PKB	protein kinase B
PKM2	pyruvate kinase M2
PMSF	phenylmethylsulfonyl fluoride
PNDM	permanent neonatal diabetes mellitus
Pre-crRNA	precursor crRNA
RAI	relative activity index
ROS	reactive oxygen species
RP	retinitis pigmentosa
S6RP	S6 ribosomal protein
SD	standard deviation

SE	standard error
SEM	standard error of the mean
sgRNA	single guide RNA
SNV	nonsynonymous substitution
STAT3	signal transducer and activator of transcription 3
TALEN	transcription activator-like effector nuclease
TCA	tricarboxylic acid
TERC	telomerase RNA component
TERT	telomerase reverse transcriptase
TIGAR	Tp53-induced glycolysis and apoptosis regulator
tracrRNA	transactivating crRNA
U	unit of enzyme's catalytic activity
VDAC	voltage-dependent anion channel
VUS	variant of uncertain significance
WB	Western blotting
ZNF	zinc-finger nuclease

REFERENCES

- Abudayyeh OO, Gootenberg JS, Essletzbichler P *et al.* RNA targeting with CRISPR-Cas13a. *Nature* (2017) **550**:280–284.
- Adzhubei IA, Schmidt S, Peshkin L *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* (2010) **7**:248–249.
- Aleshin AE, Zeng C, Bourenkov GP *et al.* The mechanism of regulation of hexokinase: new insights from the crystal structure of recombinant human brain hexokinase complexed with glucose and glucose-6-phosphate. *Structure* (1998) **6**:39–50.
- Aleshin AE, Kirby C, Liu X *et al.* Crystal structures of mutant monomeric hexokinase I reveal multiple ADP binding sites and conformational changes relevant to allosteric regulation. *J Mol Biol* (2000) **293**:1001–1015.
- Alessi DR, Andjelkovic M, Caudwell B *et al.* Mechanism of activation of protein kinase B by insulin and IGF-1. *EMBO J* (1996) **15**:6541–6551.
- Altman BJ, Hsieh AL, Sengupta A *et al.* MYC disrupts the circadian clock and metabolism in cancer cells. *Cell Metab* (2015) **22**:1009–1019.
- Amitai G & Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol* (2016) **14**:67–76.
- Anderson M, Marayati R, Moffitt R & Yeh JJ. Hexokinase 2 promotes tumor growth and metastasis by regulating lactate production in pancreatic cancer. *Oncotarget* (2017) **8**:56081–56094.
- Anderson KR, Haeussler M, Watanabe C *et al.* CRISPR off-target analysis in genetically engineered rats and mice. *Nat Methods* (2018) **15**:512–514.

Ardehali H, Printz RL, Whitesell RR *et al.* Functional interaction between the N- and C-terminal halves of human hexokinase II. *J Biol Chem* (1999) **274**:15986–15989.

Arora KK & Pedersen PL. Functional significance of mitochondrial bound hexokinase in tumor cell metabolism. Evidence for preferential phosphorylation of glucose by intramitochondrially generated ATP. *J Biol Chem* (1988) **263**:17422–17428.

Ashburner M, Ball CA, Blake JA *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* (2000) **25**:25–29.

Barnes DE. Non-homologous end joining as a mechanism of DNA repair. *Curr Biol* (2001) **11**:R455–R457.

Baysal BE, Ferrell RE, Willett-Brozick JE *et al.* Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. *Science* (2000) **287**:848–851.

Beck T & Miller BG. Structural basis for regulation of human glucokinase by glucokinase regulatory protein. *Biochemistry* (2013) **52**:6232–6239.

Bensaad K, Tsuruta A, Selak MA *et al.* TIGAR, a p53-inducible regulator of glycolysis and apoptosis. *Cell* (2006) **126**:107–120.

Bernassola F, Karin M, Ciechanover A & Melino G. The HECT family of E3 ubiquitin ligases: multiple players in cancer development. *Cancer Cell* (2008) **14**:10–21.

Bromberg Y, Kahn PC & Rost B. Neutral and weakly nonneutral sequence variants may define individuality. *Proc Natl Acad Sci USA* (2013) **110**:14255–14260.

Burnett PE, Barrow RK, Cohen NA *et al.* RAFT1 phosphorylation of the translational regulators p70 S6 kinase and 4E-BP1. *Proc Natl Acad Sci USA* (1998) **95**:1432–1437.

Bustamante E, Morris HP & Pedersen PL. High aerobic glycolysis of rat hepatoma cells in culture: role of mitochondrial hexokinase. *Proc Natl Acad Sci USA* (1977) **74**:3735–3739.

Bustamante E & Pedersen PL. Mitochondrial hexokinase of rat hepatoma cells in culture: solubilization and kinetic properties. *Biochemistry* (1980) **19**:4972–4977.

Buzzai M, Bauer DE, Jones RG *et al.* The glucose dependence of Akt-transformed cells can be reversed by pharmacologic activation of fatty acid beta-oxidation. *Oncogene* (2005) **24**:4165–4173.

Calabrese R, Capriotti E, Fariselli P *et al.* Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* (2009) **30**:1237–1244.

Canny MD, Moatti N, Wan LC *et al.* Inhibition of 53BP1 favors homology-dependent DNA repair and increases CRISPR-Cas9 genome-editing efficiency. *Nat Biotechnol* (2017) **36**:95–102.

Cantwell-Dorris ER, O’Leary JJ & Sheils OM. BRAF^{V600E}: implications for carcinogenesis and molecular therapy. *Mol Cancer Ther* (2011) **10**:385–394.

Capriotti E, Calabrese R & Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* (2006) **22**:2729–2734.

Capriotti E, Fariselli P, Rossi I & Casadio R. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* (2008) **9**:S6–S14.

Chellapa K, Jankova L, Schnabl JM *et al.* Src tyrosine kinase phosphorylation of nuclear receptor HNF4 α correlates with isoform-specific loss of HNF4 α in human colon cancer. *Proc Natl Acad Sci USA* (2012) **109**:2302–2307.

Cheung EC, Ludwig RL & Vousden KH. Mitochondrial localization of TIGAR under hypoxia stimulates HK2 and lowers ROS and cell death. *Proc Natl Acad Sci USA* (2012) **109**:20491–20496.

Christofk HR, Vander Heiden MG, Wu N *et al.* Pyruvate kinase M2 is a phosphotyrosine-binding protein. *Nature* (2008) **452**:181–186.

Christofk HR, Vander Heiden MG, Harris MH *et al.* The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* (2008) **452**:230–233.

Chow RD, Guzman CD, Wang G *et al.* AAV-mediated direct in vivo CRISPR screen identifies functional suppressors in glioblastoma. *Nat Neurosci* (2017) **20**:1329–1341.

Ciribilli Y, Singh P, Inga A & Boriak J. c-Myc targeted regulators of cell metabolism in a transgenic mouse model of papillary lung adenocarcinoma. *Oncotarget* (2016) **7**:65514–65539.

Dang CV, Kim J, Gao P & Yuste J. The interplay between MYC and HIF in cancer. *Nat Rev Cancer* (2008) **8**:51–56.

Davis EA, Cuesta-Munoz A, Raoul M *et al.* Mutants of glucokinase cause hypoglycaemia- and hyperglycaemia syndromes and their analysis illuminates fundamental quantitative concepts of glucose homeostasis. *Diabetologia* (1999) **42**:1175–1186.

DeBerardinis RJ, Lum JJ, Hatzivassiliou G & Thompson CB. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab* (2008) **7**:11–20.

Dehouck Y, Kwasigroch JM, Gilis D & Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* (2011) **12**:151–162.

Deltcheva E, Chylinski K, Sharma CM *et al.* CRISPR RNA maturation by *trans*-encoded small RNA and host factor RNase III. *Nature* (2011) **471**:602–607.

de Vooght KM, van Solinge WW, van Wesel AC *et al.* First mutation in the red blood cell-specific promoter of hexokinase combined with a novel missense mutation causes hexokinase deficiency and mild chronic hemolysis. *Haematologica* (2009) **94**:1203–1210.

DeWaal D, Nogueira V, Terry AR *et al.* Hexokinase-2 depletion inhibits glycolysis and induces phosphorylation in hepatocellular carcinoma and sensitizes to metformin. *Nat Commun* (2018) **9**:446.

Doench JG, Fusi N, Sullender M *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR/Cas9. *Nat Biotechnol* (2016) **34**:184–191.

Fang R, Xiao T, Fang Z *et al.* MicroRNA-143 (miR-143) regulates cancer glycolysis via targeting hexokinase 2 gene. *J Biol Chem* (2012) **287**:23227–23235.

Fantin VR, St-Pierre J & Leder P. Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell* (2006) **9**:425–434.

Fernandez PC, Frank SR, Wang L *et al.* Genomic targets of the human c-Myc protein. *Genes Dev* (2003) **17**:1115–1129.

Feron O. Pyruvate into lactate and back: from the Warburg effect to symbiotic energy fuel exchange in cancer cells. *Radiother Oncol* (2009) **92**:329–333.

Flanagan SE, Patch AM & Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers* (2010) **14**:533–537.

Fonfara I, Richter H, Bratovic M *et al.* The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* (2016) **532**:517–521.

Fradet-Turcotte A, Canny MD, Escribano-Diaz C *et al.* 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature* (2013) **499**:50–54.

Fu Y, Foden JA, Khayter C, Maeder ML *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* (2013) **31**:822–826.

Gao P, Tchernyshyov I, Chang TC *et al.* c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature* (2009) **458**:462–765.

García-Herrero CM, Galán M, Vincent O *et al.* Functional analysis of human glucokinase gene mutations causing MODY2: exploring the regulatory mechanisms of glucokinase activity. *Diabetologia* (2007) **50**:325–333.

Gloyn AL. Glucokinase (*GCK*) mutations in hyper- and hypoglycemia: maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemia of infancy. *Hum Mutat* (2003) **22**:353–362.

Gloyn AL, Odili S, Zelent D *et al.* Insights into the structure and regulation of glucokinase from a novel mutation (V62M), which causes maturity-onset diabetes of the young. *J Biol Chem* (2005) **280**:14105–14113.

Graham NA, Minasyan A, Lomova A *et al.* Recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures. *Mol Syst Biol* (2017) **13**:914.

Green MR & Sambrook J. *Molecular cloning: a laboratory manual* (fourth edition). Cold Spring Harbor Laboratory Press, New York; (2012).

Gregersen LH, Jacobsen A, Frankel LB *et al.* MicroRNA-143 down-regulates hexokinase 2 in colon cancer cells. *BMC Cancer* (2012) **12**:232.

Gupta A, Ajith A, Singh S *et al.* PAK2-c-Myc-PKM2 axis plays an essential role in head and neck oncogenesis via regulating Warburg effect. *Cell Death Dis* (2018) **9**:825.

Hale CR, Zhao P, Olson S *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* (2009) **139**:945–956.

Hardie DG. AMP-activated/SNF1 protein kinases: conserved guardians of cellular energy. *Nat Rev Mol Cell Biol* (2007) **8**:774–785.

Haurwitz RE, Jinek M, Wiedenheft B *et al.* Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* (2010) **329**:1355–1358.

Hecht M, Bromberg Y & Rost B. Better prediction of functional effects for sequence variants. *BMC Genomics* (2015) **16**:S1–S12.

Hockemeyer D, Wang H, Kiani S *et al.* Genetic engineering of human pluripotent cells using TALE nucleases. *Nat Biotechnol* (2011) **29**:731–734.

Hopf TA, Ingraham JB, Poelwijk FJ *et al.* Mutation effects predicted from sequence co-variation. *Nat Biotechnol* (2017) **33**:128–135.

Houllberghs H, Dekker M, Lantermans H *et al.* Oligonucleotide-directed mutagenesis screen to identify pathogenic Lynch syndrome-associated MSH2 DNA mismatch repair gene variants. *Proc Natl Acad Sci USA* (2016) **113**:4128–4133.

Hustedt N & Durocher D. The control of DNA repair by the cell cycle. *Nat Cell Biol* (2016) **19**:1–9.

Ihry RJ, Worringer KA, Salick MR *et al.* p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nat Med* (2018) **24**:939–946.

- Ioannidis NM, Rothstein JH, Pejaver V *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* (2016) **99**:877–885.
- Jacinto E, Loewith R, Schmidt A *et al.* Mammalian TOR complex 2 controls the actin cytoskeleton and is rapamycin insensitive. *Nat Cell Biol* (2004) **6**:1122–1128.
- Jackson RN, Golden SM, van Erp PB *et al.* Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* (2014) **345**:1473–1479.
- Jain A, Zode G, Kasetti RB *et al.* CRISPR-Cas9-based treatment of myocilin-associated glaucoma. *Proc Natl Acad Sci USA* (2017) **114**:11199–11204.
- Jerath NU & Shy ME. Hereditary motor and sensory neuropathies: Understanding molecular pathogenesis could lead to future treatment strategies. *Biochim Biophys Acta* (2015) **1852**:667–678.
- Jiang S, Zhang LF, Zhang HW *et al.* A novel miR-155/miR-143 cascade controls glycolysis by regulating *hexokinase 2* in breast cancer cells. *EMBO J* (2012) **31**:1985–1998.
- Jiao L, Zhang HL, Li DD *et al.* Regulation of glycolytic metabolism by autophagy in liver cancer involves selective autophagic degradation of HK2 (hexokinase 2). *Autophagy* (2018) **14**:671–684.
- Jinek M, Chylinski K, Fonfara I *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* (2012) **337**:816–821.
- Jinek M, Jiang F, Taylor DW *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* (2014) **343**:1247997.
- Johansen CT, Wang J & Lanktree MB. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet* (2010) **42**:684–687.

John S, Weiss JN & Ribalet B. Subcellular localization of hexokinases I and II directs the metabolic fate of glucose. *PLoS ONE* (2011) **6**:e17674.

Joung J, Konermann S, Gootenberg JS *et al.* Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat Protoc* (2017) **13**:828–863.

Julien LA, Carriere A, Moreau J & Roux PP. mTORC1-activated S6K1 phosphorylates Rictor on threonine 1135 and regulates mTORC2 signaling. *Mol Cell Biol* (2010) **30**:908–921.

Kamata K, Mitsuya M, Nishimura T *et al.* Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure* (2004) **12**:429–438.

Kato S, Han SY, Liu W *et al.* Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci USA* (2003) **100**:8424–8429.

Kazyken D, Magnuson B, Bodur C *et al.* AMPK directly activates mTORC2 to promote cell survival during acute energetic stress. *Sci Signal* (2019) **12**:eaav3249.

Kennedy KM & Dewhirst MW. Tumor metabolism of lactate: the influence and therapeutic potential for MCT and CD147 regulation. *Future Oncol* (2010) **6**:127–148.

Kim DH, Sarbassov DD, Ali SM *et al.* mTOR interacts with raptor to form a nutrient-sensitive complex that signals to the cell growth machinery. *Cell* (2002) **110**:163–175.

Kim DH, Sarbassov DD, Ali SM *et al.* GbetaL, a positive regulator of the rapamycin-sensitive pathway required for the nutrient-sensitive interaction between raptor and mTOR. *Mol Cell* (2003) **11**:895–904.

Kim JW, Gao P, Liu YC *et al.* Hypoxia-inducible factor 1 and dysregulated c-Myc cooperatively induce vascular endothelial growth factor and metabolic switches hexokinase 2 and pyruvate dehydrogenase kinase 1. *Mol Cell Biol* (2007) **27**:7381–7393.

Kim D, Kim J, Hur JK *et al.* Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat Biotechnol* (2016) **34**:863–868.

King A, Selak MA & Gottlieb E. Succinate dehydrogenase and fumarate hydratase: linking mitochondrial dysfunction and cancer. *Oncogene* (2006) **25**:4675–4682.

Kleinvister BP, Tsai SQ, Prew MS *et al.* Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat Biotechnol* (2016) **34**:869–874.

Koonin EV, Makarova KS & Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* (2017) **37**:67–78.

Kosicki M, Tomberg K & Bradley A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat Biotechnol* (2018) **36**:765–771.

Kudryavtseva AV, Fedorova MS, Zhanoronkov A *et al.* Effect of lentivirus-mediated shRNA inactivation of HK1, HK2, and HK3 genes in colorectal cancer and melanoma cells. *BMC Genet* (2016) **17**:156.

Kumar MD, Bava KA, Gromiha MM *et al.* ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res* (2006) **34**:D204–D206.

Kumar P, Henikoff S & Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* (2009) **4**:1073–1081.

Kunin V, Sorek R & Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* (2007) **8**:R61.

Labun K, Montague TG, Gagnon JA *et al.* CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res* (2016) **44**:W272–W276.

Labun K, Montague TG, Krause M *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res* (2019) **47**:W171–W174.

Lee HJ, Li CF, Ruan D *et al.* Non-proteolytic ubiquitination of hexokinase 2 by HectH9 controls tumor metabolism and cancer stem cell expansion. *Nat Commun* (2019) **10**:2625.

Li F, Wang Y, Zeller K *et al.* Myc stimulates nuclearly encoded mitochondrial genes and mitochondrial biogenesis. *Mol Cell Biol* (2005) **25**:6225–6234.

Li T, Guo H, Song Y *et al.* Loss of vinculin and membrane-bound β -catenin promotes metastasis and predicts poor prognosis in colorectal cancer. *Mol Cancer* (2014) **13**:263.

Liang Y, Kesavan P, Wang L *et al.* Variable effects of maturity-onset-diabetes-of youth (MODY)-associated glucokinase mutations on substrate interactions and stability of the enzyme. *Biochem J* (1995) **309**:167–173.

Liu L, Muralidhar S, Singh M *et al.* High-density SNP genotyping to define beta-globin locus haplotypes. *Blood Cells Mol Dis* (2009) **42**:16–24.

Liu JJ, Orlova N, Oakes BL *et al.* CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* (2019) **566**:218–223.

Majmundar AJ, Wong WJ & Simon MC. Hypoxia-inducible factors and the response to hypoxic stress. *Mol Cell* (2010) **40**:294–309.

Makarova KS, Haft DH, Barrangou R *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* (2011) **9**:467–477.

Martin PL, Yin JJ, Seng V *et al.* Androgen deprivation leads to increased carbohydrate metabolism and hexokinase 2-mediated survival in *Pten/Tp53*-deficient prostate cancer. *Oncogene* (2017) **36**:525–533.

Masui K, Tanaka K, Akhavan D *et al.* mTOR complex 2 controls glycolytic metabolism in glioblastoma through FoxO acetylation and upregulation of c-Myc. *Cell Metab* (2013) **18**:726–739.

Mathe E, Olivier M, Kato S *et al.* Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res* (2006) **34**:1317–1325.

Matschinsky FM, Davis EA, Cuesta-Munoz A *et al.* The glucokinase system and the regulation of blood sugar. In: Matschinsky DM, Magnuson MA (eds) *Molecular pathogenesis of MODYs*. Karger, Basel, (2000) pp 99–108.

Matschinsky FM. Assessing the potential of glucokinase activators in diabetes therapy. *Nat Rev Drug Discov* (2009) **8**:399–416.

Maxwell KN, Hart SN, Vijai J *et al.* Evaluation of ACMG-guideline based variant classification of cancer susceptibility and non-cancer-associated genes in families affected by breast cancer. *Am J Hum Genet* (2016) **98**:801–817.

Mergenthaler P, Kahl A, Kamitz A *et al.* Mitochondrial hexokinase II (HKII) and phosphoprotein enriched in astrocytes (PEA15) form a molecular switch governing cellular fate depending on the metabolic state. *Proc Natl Acad Sci USA* (2012) **109**:1518–1523.

Milenković T, Zdravković D & Mitrović K. [Novel glucokinase mutation in a boy with maturity-onset diabetes of the young]. *Srp Arh Celok Lek* (2008) **136**:542–544.

Miller JC, Holmes MC, Wang J *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat Biotechnol* (2007) **25**:778–785.

Miller JC, Tan S, Qiao G *et al.* A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* (2011) **29**:143–148.

Mojica FJ, Diez-Villasenor C, Garcia-Martinez J & Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defense system. *Microbiology* (2009) **166**:733–740.

Moreno-Sanchez R, Rodriguez-Enriquez S, Marin-Hernandez A & Saavedra E. Energy metabolism in tumor cells. *FEBS J* (2007) **274**:1393–1418.

Morrish F, Noonan J, Perez-Olsen C *et al.* Myc-dependent mitochondrial generation of acetyl-CoA contributes to fatty acid biosynthesis and histone acetylation during cell cycle entry. *J Biol Chem* (2010) **285**:36267–36274.

Myhrvold C, Freije CA, Gootenberg JS *et al.* Field-deployable viral diagnostics using CRISPR-Cas13. *Science* (2018) **360**:444–448.

Najm FJ, Strand C, Donovan KF *et al.* Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. *Nat Biotechnol* (2018) **36**:179–189.

Nakashima RA, Mangan PS, Colombini M & Pedersen PL. Hexokinase receptor complex in hepatoma mitochondria: evidence from N,N'-dicyclohexylcarbodiimide-labeling studies for the involvement of the pore-forming protein VDAC. *Biochemistry* (1986) **25**:1015–1021.

Nykamp K, Anderson M, Powers M *et al.* Sherlock: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med* (2017) **19**:1105–1117.

Orthwein A, Noordermeer SM, Wilson MD *et al.* A mechanism for the suppression of homologous recombination in G1 cells. *Nature* (2015) **528**:422–426.

Osbak KK, Colclough K, Saint-Martin C *et al.* Update on mutations in glucokinase (*GCK*), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. *Hum Mutat* (2009) **30**:1512–1526.

Osthus RC, Shim H, Kim S *et al.* Deregulation of glucose transporter 1 and glycolytic gene expression by c-Myc. *J Biol Chem* (2000) **275**:21797–21800.

Paglia DE, Shende A, Lanzkowsky P & Valentine WN. Hexokinase “New Hyde Park”: A low activity erythrocyte isozyme in a Chinese kindred. *Am J Hematol* (1981) **10**:107.

Parsons DW, Jones S, Zhang X *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* (2008) **321**:1807–1812.

Peck B, Ferber EC & Schulze A. Antagonism between FOXO and MYC regulates cellular powerhouse. *Front Oncol* (2013) **3**:96.

Pedersen PL. Tumor mitochondria and the bioenergetics of cancer cells. *Prog Exp Tumor Res* (1978) **22**:190–274.

Pinterova D, Ek J, Kolostova K *et al.* Six novel mutations in the *GCK* gene in MODY patients. *Clin Genet* (2007) **71**:95–96.

Pruhova S, Dusatkova P, Sumnik Z *et al.* Glucokinase diabetes in 103 families from a country-based study in the Czech Republic: geographically restricted distribution of two prevalent *GCK* mutations. *Pediatr Diabetes* (2010) **11**:529–535.

Qi LS, Larson MH, Gilbert LA *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* (2013) **152**:1173–1183.

Ran FA, Hsu PD, Wright J *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* (2013) **8**:2281–2308.

Ran FA, Hsu PD, Lin CY *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* (2013) **154**:1380–1389.

Rees MG, Ng D, Ruppert S *et al.* Correlation of rare coding variants in the gene encoding human glucokinase regulatory protein with phenotypic, cellular, and kinetic outcomes. *J Clin Invest* (2012) **122**:205–217.

Richards S, Aziz N, Bale S *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* (2015) **17**:405–424.

Rijksen G, Akkerman JW, van den Wall Bake AW *et al.* Generalized hexokinase deficiency in the blood cells of a patient with nonspherocytic hemolytic anemia. *Blood* (1983) **61**:12–18.

Rizzo MA & Piston DW. Regulation of β cell glucokinase by S-nitrosylation and association with nitric oxide synthase. *J Cell Biol* (2003) **161**:243–248.

Roh J, Kim Y, Oh J *et al.* Hexokinase 2 is a molecular bridge linking telomerase and autophagy. *PLoS ONE* (2018) **13**:e0193182.

Romeo S, Yin W, Kozlitina J *et al.* Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J Clin Invest* (2009) **119**:70–79.

Ruvinsky I, Sharon N, Lerer T *et al.* Ribosomal protein S6 phosphorylation is a determinant of cell size and glucose homeostasis. *Genes Dev* (2005) **19**:2199–2211.

Sarbassov DD, Guertin DA, Ali SM & Sabatini DM. Phosphorylation and regulation of Akt/PKB by the rictor-mTOR complex. *Science* (2005) **307**:1098–1101.

Sagen JV, Odili S, Bjorkhaug L *et al.* From clinicogenetic studies of maturity-onset diabetes of the young to unraveling complex mechanisms of glucokinase regulation. *Diabetes* (2006) **55**:1713–1722.

Saleh-Gohari N & Helleday T. Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells. *Nucleic Acids Res* (2004) **32**:3683–3688.

Sapranauskas R, Gasiunas G, Fremaux C *et al.* The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res* (2011) **39**:9275–9282.

Saxton RA & Sabatini DM. mTOR signaling in growth, metabolism, and disease. *Cell* (2017) **168**:960–976.

Schmid-Burgk JL, Schmidt T, Kaiser V *et al.* A ligation-independent cloning technique for high-throughput assembly of transcription activator-like effector genes. *Nat Biotechnol* (2013) **31**:76–81.

Schunder E, Rydzemski K, Grunow R & Heuner K. First indication for a functional CRISPR/Cas system in *Francisella tularensis*. *Int J Med Microbiol* (2013) **303**:51–60.

Schwank G, Koo BK, Sasselli V *et al.* Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients. *Cell Stem Cell* (2013) **13**:653–658.

Scott DA & Zhang F. Implications of human genetic variation on CRISPR-based therapeutic genome editing. *Nat Med* (2017) **23**:1095–1101.

Shim H, Dolde C, Lewis BC *et al.* c-Myc transactivation of LDH-A: implications for tumor metabolism and growth. *Proc Natl Acad Sci USA* (1997) **94**:6658–6663.

Simcikova D, Kockova L, Vackarova K *et al.* Evidence-based tailoring of bioinformatics approaches to optimize methods that predict the effects of nonsynonymous amino acid substitutions in glucokinase. *Sci Rep* (2017) **7**:9499.

Simcikova D & Heneberg P. Refinement of evolutionary medicine predictions based on clinical evidence for the manifestations of Mendelian diseases. Under revision in *Sci Rep* (subm.) (2019).

Steele AM, Tribble ND, Caswell R *et al.* The previously reported T342P GCK missense variant is not a pathogenic mutation causing MODY. *Diabetologia* (2011) **54**:2202–2205.

Stenson PD, Mort M, Ball EV *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* (2014) **133**:1–9.

Sullivan LS, Koboldt DC, Bowne SJ *et al.* A dominant mutation in hexokinase 1 (*HK1*) causes retinitis pigmentosa. *Invest Ophthalmol Vis Sci* (2014) **55**:7147–7158.

Tavtigian SV, Deffenbaugh AM, Yin L *et al.* Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* (2005) **43**:295–305.

Thakore PI, D'Ippolito AM, Song L *et al.* Highly specific epigenome editing by CRISPR/Cas9 repressors for silencing of distal regulatory elements. *Nat Methods* (2015) **12**:1143–1149.

The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res* (2015) **43**:D1049–D1056.

Treins C, Warne PH, Magnuson MA *et al.* Rictor is a novel target of p70 S6 kinase-1. *Oncogene* (2010) **29**:1003–1016.

Tsai HJ & Wilson JE. Functional organization of mammalian hexokinases: characterization of the rat type III isozyme and its chimeric forms, constructed with the N- and C-terminal halves of the type I and type II isozymes. *Arch Biochem Biophys* (1997) **338**:183–192.

Van den Bosch M, Lohman PH & Pastink A. DNA double-strand break repair by homologous recombination. *Biol Chem* (2002) **383**:873–892.

Vander Heiden MG, Chandel NS, Schumacker PT & Thompson CB. Bcl-xL prevents cell death following growth factor withdrawal by facilitating mitochondrial ATP/ADP exchange. *Mol Cell* (1999) **3**:159–167.

Vander Heiden MG, Cantley LC & Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* (2009) **324**:1029–1033.

Walsh R, Thomson KL, Ware JS *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med* (2016) **19**:192–203.

Wang F, Wang Y, Zhang B *et al.* A missense mutation in *HK1* leads to autosomal dominant retinitis pigmentosa. *Invest Ophthalmol Vis Sci* (2014) **55**:7159–7164.

Warburg O & Dickens F. Kaiser Wilhelm-Institut für Biologie B. The metabolism of tumours: investigations from the Kaiser-Wilhelm Institute for Biology, London: Constable. Berlin-Dahlem; 1930.

Whittington AC, Larion M, Bowler JM *et al.* Dual allosteric activation mechanisms in monomeric human glucokinase. *Proc Natl Acad Sci USA* (2015) **112**:11553–11558.

Wiedenheft B, Sternberg SH & Doudna JA. RNA-guided genetic silencing systems in bacteria and archae. *Nature* (2012) **482**:331–338.

Wilson J.E. Isozymes of mammalian hexokinase: structure, subcellular localization and metabolic function. *J Exp Biol* (2003) **206**:2049–2057.

Wood AJ, Lo TW, Zeitler B *et al.* Targeted genome editing across species using ZFNs and TALENs. *Science* (2011) **333**:307.

Wu J, Hu L, Wu F *et al.* Poor prognosis of hexokinase 2 overexpression in solid tumors of digestive system: a meta-analysis. *Oncotarget* (2017) **8**:32332–32344.

Xue W, Chen S, Yin H *et al.* CRISPR-mediated direct mutation of cancer genes in the mouse liver. *Nature* (2014) **514**:380–384.

Yamano T, Nishimasu H, Zetsche B *et al.* Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell* (2016) **165**:949–962.

Yap, CS, Peterson AL, Castellani G *et al.* Kinetic profiling of the c-Myc transcriptome and bioinformatic analysis of repressed gene promoters. *Cell Cycle* (2011) **10**:2184–2196.

Yin H, Xue W, Chen S *et al.* Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. *Nat Biotechnol* (2014) **32**:551–553.

Zetsche B, Gootenberg JS, Abudayyeh OO *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* (2015) **163**:759–771.

Zhang F, Cong L, Lodato S *et al.* Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotechnol* (2011) **29**:149–153.

Zhang J, Wang S, Jiang B *et al.* c-Src phosphorylation and activation of hexokinase promotes tumorigenesis and metastasis. *Nat Commun* (2017) **8**:13732.

Zhao H, Sheng G, Wang J *et al.* Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. *Nature* (2014) **515**:147–150.