



REVIEW-SYMPOSIUM

Pitfalls of using sequence databases for heterologous expression studies – a technical review

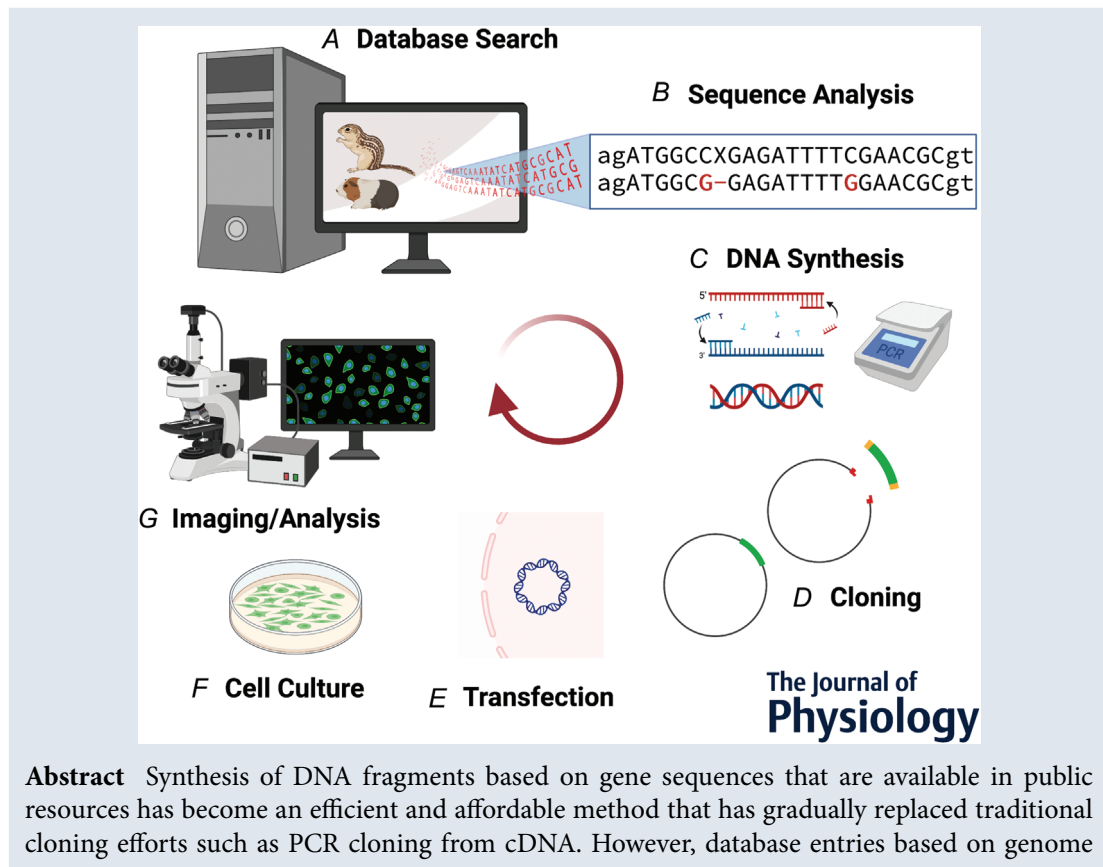
Stephan Maxeiner¹ , Gabriela Krasteva-Christ¹  and Mike Althaus²

¹Institute for Anatomy and Cell Biology, Saarland University, Homburg, Germany

²Department of Natural Sciences, Institute for Functional Gene Analytics, Bonn-Rhein-Sieg University of Applied Sciences, Rheinbach, Germany

Handling Editors: Laura Bennet & Peking Fong

The peer review history is available in the Supporting information section of this article (<https://doi.org/10.1113/JP284066#support-information-section>).



Stephan Maxeiner is a research associate in the Department of Anatomy at Saarland University, Homburg, Germany. He received his training as a PhD student at Bonn University, Germany, working on gap junctional communication in the lab of Klaus Willecke. As a postdoctoral researcher, he joined the lab of Thomas Südhof at UT Southwestern, Dallas, Texas and Stanford University, California, USA, where he studied mouse models of neurodevelopmental diseases such as autism spectrum disorders. Stephan is currently interested in human genes with clinical pathologies that are absent in classical rodent research animals such as mice and rats. **Mike Althaus** trained at the University of Giessen and is now a Professor in Physiology at the Bonn-Rhein-Sieg University of Applied Sciences in Germany. He leads an Ion Transport Research Group and investigates the molecular physiology of ion channels and transporters in health and disease.



sequencing results are prone to errors which can lead to false sequence information and, ultimately, errors in functional characterisation of proteins such as ion channels and transporters in heterologous expression systems. We have identified five common problems that repeatedly appear in public resources: (1) Not every gene has yet been annotated; (2) not all gene annotations are necessarily correct; (3) transcripts may contain automated corrections; (4) there are mismatches between gene, mRNA and protein sequences; and (5) splicing patterns often lack experimental validation. This technical review highlights and provides a strategy to bypass these issues in order to avoid critical mistakes that could impact future studies of any gene/protein of interest in heterologous expression systems.

(Received 7 December 2022; accepted after revision 7 February 2023; first published online 9 February 2023)

Corresponding author Stephan Maxeiner: Institute for Anatomy and Cell Biology, Saarland University, Homburg, Germany. Email: stephan.maxeiner@uni-saarland.de

Abstract figure legend Projects involving heterologous gene expression are often characterised by similar steps. Initially, database research (A) is necessary to retrieve information of full or partial sequences of a gene of interest. A multitude of genome assemblies are annotated and deposited in public databases or are available for refined search options using individual sequence information. The search results need to be scrutinised and compared with already available information (B). Once the sequence has been determined, DNA synthesis (C) by PCR or commercial synthesis is necessary for further cloning procedures (D). Eventually, the DNA needs to be transfected (E) and expressed in, for example, eukaryotic cells (F). Finally, the expression of the gene of interest needs to be documented and its function analysed (G).

Introduction

Heterologous overexpression systems such as *Xenopus laevis* oocytes, human embryonic kidney 293 (HEK-293) cells or Chinese hamster ovary (CHO) cells are commonly employed models for the functional characterisation of ion channels and transporters with various electrophysiological techniques, including structure–function analyses or pharmacological research (Ooi et al., 2016; Papke & Smith-Maxwell, 2009). In order to express membrane proteins in *Xenopus* oocytes, RNA transcripts of cDNA (cRNA), coding for the open reading frame (ORF) of the protein of interest, are commonly injected into the cytoplasm of the oocytes. This yields efficient translation and incorporation of the protein into the plasma membrane (Bhatt et al., 2022). HEK-293 or CHO cells are usually transiently or permanently transfected with cDNA encoding the ORF of the protein of interest. Consequently, both techniques require the precise genetic information for the protein to be studied. Historically, most labs generated cDNA fragments encoding the ORF of the desired ion channel or transporter by cloning it from cDNA libraries which were obtained following mRNA isolation from cells or tissues (e.g. Fronius et al., 2010). These cDNA sequences, therefore, corresponded to mRNA transcripts which were endogenously present in the cells or tissues of the organism from which the isolate was derived. Due to technological advances over the past decade, synthesis of DNA fragments based on gene sequences that are available in public resources, for example the National Centre for Biotechnology

Information (NCBI, Bethesda, MD, USA), has become an efficient and affordable method that has gradually replaced traditional cloning efforts such as PCR cloning from cDNA. Furthermore, the growing number of available genomes from different species paves the way for comparative molecular biology and physiology. For example, the number of published rodent genomes has expanded dramatically within recent years, helping to assess the evolutionary fate of genes on different suborder branches (Fig. 1). However, while traditional cloning strategies ensured that obtained cDNA sequences represented naturally occurring transcripts in an organism, database entries based on genome sequencing efforts are prone to errors which can lead to false sequence information and, ultimately, the functional characterisation of proteins, for example ion channels or transporters, that do not exist naturally in a given species. In addition to the use of public resources in overexpression studies, RNA sequencing (RNA-Seq) has emerged as a powerful tool for gene expression studies. Ultimately, this technique requires the comparison of the obtained sequences with database entries in order to identify the protein which is encoded by the RNA. Mistakes in database entries, for example the lack of annotation of a specific gene, will therefore impact the interpretation of RNA-Seq results as well.

Frequently encountered problems

In our previous investigation on epithelial sodium channel genes, *SCNN1* (Gettings et al., 2021; Wichmann et al.,

2019), as well as efforts to identify mammalian research models for genes of the pseudoautosomal region (PAR) which have clinical implications (Maxeiner et al., 2021), we repeatedly encountered problems with sequence data from public resources. For example, studies involving the evolutionary fate of PAR genes often demand knowledge of the biological sex of the DNA donor, since the PAR boundary, that is, the boundary between PAR and the X-specific part of the X-chromosome as well as the male-specific part of the Y-chromosome, is showing transitional changes (Maxeiner et al., 2021). Additionally, knowledge of differences in gametologues, that is, genes on the X- and Y-chromosomes that share the same ancestor, for example *AMELX/Y* (Akane et al., 1991) or *NLGN4X/Y*, is useful in order to develop sex-typing strategies (Maxeiner et al., 2019, 2022; Zaffalon et al., 2019). This is only possible if ideally both, that is, male/female genomes or at least the male genome have been fully sequenced. However, in the case of rodents, there is no biological sex specified by the submitting party in almost one third of the published genomes (Fig. 1B). This does not mean that X-chromosomal genes are generally missing from the respective assemblies. It is, however, difficult to assess whether the failure to detect, for example, the *SRY* gene (*sex determining region on Y*), a marker gene present on the Y-chromosome, is due to its constitutive absence from a supposedly female genome or its pronounced sequence divergence escaping sequence alignment tools. In addition to the lack of biological sex assignment, we identified five categories of problem with the use and interpretation of data from public resources. In the following sections, we will outline each problem with specific examples and address its origin. We suggest a strategy to bypass

these issues in order to avoid critical mistakes that could impact future studies of any gene/protein of interest in heterologous expression systems. Observations reported herein resulted from detailed studies of the above-mentioned genes of interest and should not be perceived as a result of in-depth whole genome comparisons.

Problem 1 – Not every gene has yet been annotated

In the process of assessing critical sequence information for any subsequent study, a distinction between the availability of genomic information and the annotation of a given genome needs to be made. Annotated genomes are found, for instance, in the 'gene' section of the National Centre of Biotechnology Information (NCBI), hosted by the US National Institutes of Health (NIH; <https://www.ncbi.nlm.nih.gov/gene/>). As a rule of thumb, housekeeping genes, that is, genes that are important for maintaining critical cellular or physiological functions, appear to retain very similar sequences across a broad range of species. These genes generally appear properly annotated. Genes that are exclusive to certain vertebrate animal taxa (fish, amphibians, reptiles/birds and mammals) or even subdivisions thereof, might not be fully annotated despite the publication of the species genome. For example, this is the case for the *SCNN1D* genes which code for the δ -subunit of the epithelial sodium channel (ENaC) in vertebrates. Within the group of annotated rodent genomes (Table 1), there are currently only 18 entries for *SCNN1D* as compared with the housekeeping gene beta-actin, *ACTB*, listing 31 entries (<https://www.ncbi.nlm.nih.gov/gene/>; accessed on 16.10.2022 by

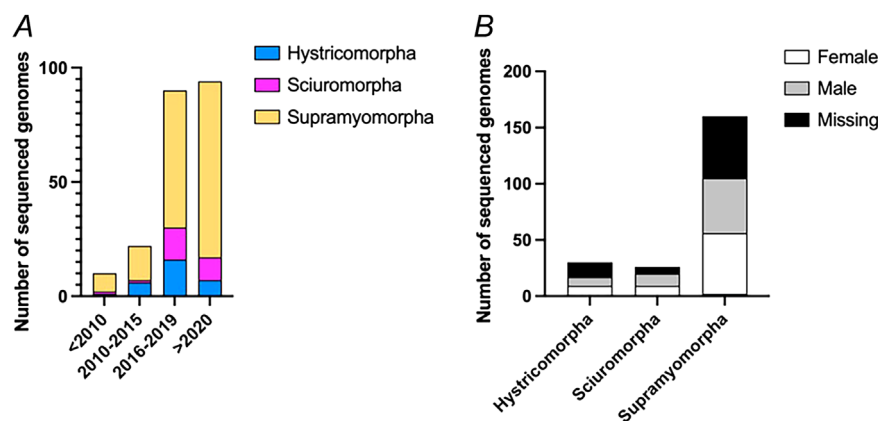


Figure 1. Current and historical development of the sequencing status of rodent genomes

A, during the last two decades the largest increase in the number of sequenced genomes occurred after 2015. The suborder Supramyomorpha (which includes the family of Muridae (mouse-like species)) contributed the largest addition to the number of sequenced genomes. B, the biological sex has not been assigned in a considerable fraction of sequenced genomes. Samples derived from female specimens potentially impact research related to the pseudoautosomal region and Y-chromosome evolution.

Table 1. List of sequenced rodent genomes

Sciuromorpha	
Aplodontidae	<i>Aplodontia rufa</i> (mountain beaver)
Gliridae	<i>Glis glis</i> (fat dormouse), <i>Graphiurus murinus</i> (woodland dormouse), <i>Muscardinus avellanarius</i> (hazel dormouse)
Sciuridae	<i>Cynomys gunnisoni</i> (Gunnison's prairie dog), <i>Ictidomys tridecemlineatus</i> ¹ (thirteen-lined ground squirrel), <i>Glaucomys volans</i> (southern flying squirrel), <i>Marmota flaviventris</i> ¹ (yellow-bellied marmot), <i>Marmota himalayana</i> (Himalayan marmot), <i>Marmota marmota marmota</i> ¹ (Alpine marmot), <i>Marmota monax</i> ¹ (woodchuck), <i>Marmota vancouverensis</i> (Vancouver Island marmot), <i>Neosciurus carolinensis</i> (grey squirrel), <i>Sciurus niger</i> (fox squirrel), <i>Sciurus vulgaris</i> (Eurasian red squirrel), <i>Spermophilus dauricus</i> (Daurian ground squirrel), <i>Urocitellus parryii</i> ¹ (Arctic ground squirrel), <i>Xerus inauris</i> (South African ground squirrel)
Supramyomorpha	
Castoridae	<i>Castor canadensis</i> ¹ (American beaver)
Cricetidae	<i>Arvicola amphibius</i> ¹ (Eurasian water vole), <i>Cricetulus griseus</i> ¹ (Chinese hamster), <i>Ellobius lutescens</i> (Transcaucasian mole vole), <i>Ellobius talpinus</i> (northern mole vole), <i>Mesocricetus auratus</i> ¹ (golden hamster), <i>Microtus agrestis</i> (short-tailed field vole), <i>Microtus arvalis</i> (common vole), <i>Microtus fortis</i> (reed vole), <i>Microtus montanus</i> (montane vole), <i>Microtus ochrogaster</i> ¹ (prairie vole), <i>Microtus oeconomus</i> (root vole), <i>Microtus oregoni</i> (creeping vole), <i>Microtus richardsoni</i> (water vole), <i>Myodes glareolus</i> ¹ (bank vole), <i>Neodon shergylaensis</i> ² , <i>Neotoma lepida</i> (desert woodrat), <i>Ondatra zibethicus</i> (muskrat), <i>Onychomys torridus</i> ¹ (southern grasshopper mouse), <i>Peromyscus attwateri</i> (Texas deer mouse), <i>Peromyscus aztecus</i> (Aztec mouse), <i>Peromyscus californicus insignis</i> (California mouse), <i>Peromyscus eremicus</i> (cactus mouse), <i>Peromyscus leucopus</i> ¹ (white-footed mouse), <i>Peromyscus maniculatus bairdii</i> ¹ (prairie deer mouse), <i>Peromyscus melanophrys</i> (plateau mouse), <i>Peromyscus nudipes</i> ² , <i>Peromyscus polionotus subgriseus</i> (oldfield mouse), <i>Phodopus roborovskii</i> (desert hamster), <i>Phodopus sungorus</i> ^{1,2} , <i>Sigmodon hispidus</i> (hispid cotton rat)
Dipodidae	<i>Jaculus jaculus</i> ¹ (lesser Egyptian jerboa), <i>Orientalactaga bullata</i> (Gobi jerboa)
Geomyidae	<i>Thomomys bottae</i> (Botta's pocket gopher)
Heteromyidae	<i>Dipodomys merriami</i> (Merriam's kangaroo rat), <i>Dipodomys ordii</i> ¹ (Ord's kangaroo rat), <i>Dipodomys spectabilis</i> ¹ (banner-tailed kangaroo rat), <i>Dipodomys stephensi</i> (Stephens's kangaroo rat), <i>Perognathus longimembris pacificus</i> ¹ (Pacific pocket mouse)
Muridae	<i>Acomys cahirinus</i> (Egyptian spiny mouse), <i>Acomys dimidiatus</i> ² , <i>Acomys kempii</i> (Kemp's spiny mouse), <i>Acomys percivali</i> (Percival's spiny mouse), <i>Acomys russatus</i> (golden spiny mouse), <i>Apodemus speciosus</i> (large Japanese field mouse), <i>Apodemus sylvaticus</i> (European woodmouse), <i>Arvicanthis niloticus</i> ¹ (African grass rat), <i>Grammomys dolichurus</i> (common thicket rat), <i>Grammomys surdaster</i> ^{1,2} , <i>Hylomyscus alleni</i> (Allen's wood mouse), <i>Lophiomys imhausi</i> (crested rat), <i>Mastomys coucha</i> ¹ (southern multimammate mouse), <i>Mastomys natalensis</i> (African soft-furred rat), <i>Meriones unguiculatus</i> ¹ (Mongolian gerbil), <i>Mus caroli</i> ¹ (Ryukyu mouse), <i>Mus minutoides</i> (Southern African pygmy mouse), <i>Mus musculus</i> ¹ (house mouse), <i>Mus musculus castaneus</i> ¹ (southeastern Asian house mouse), <i>Mus musculus domesticus</i> (western European house mouse), <i>Mus musculus molossinus</i> (Japanese wild mouse), <i>Mus pahari</i> ¹ (shrew mouse), <i>Mus spicilegus</i> (steppe mouse), <i>Mus spretus</i> (western wild mouse), <i>Praomys delectorum</i> (delectable soft-furred mouse), <i>Psammomys obesus</i> (fat sand rat), <i>Rattus norvegicus</i> ¹ (Norway rat), <i>Rattus rattus</i> ¹ (black rat), <i>Rhabdomys dilectus</i> (mesic four-striped grass rat), <i>Rhombomys opimus</i> (great gerbil), <i>Rhynchomys soricoides</i> (Mount Data shrew rat)
Nesomyidae	<i>Cricetomys gambianus</i> (Gambian giant pouched rat)
Pedetidae	<i>Pedetes capensis</i> (springhare)
Platacanthomyidae	<i>Typhlomys cinereus</i> (soft-furred tree mouse)
Spalacidae	<i>Nannospalax galili</i> ¹ (Upper Galilee mountains blind mole-rat), <i>Rhizomys pruinosus</i> (hoary bamboo rat)
Zapodidae	<i>Zapus hudsonius</i> (meadow jumping mouse)
Hystricomorpha	
Bathyergidae	<i>Fukomys damarensis</i> ¹ (Damara mole-rat)
Caviidae	<i>Cavia porcellus</i> ¹ (domestic guinea pig), <i>Cavia aperea</i> (Brazilian guinea pig), <i>Cavia tschudii</i> (Montane guinea pig), <i>Dolichotis patagonum</i> (Patagonian cavy), <i>Hydrochoerus hydrochaeris</i> (capybara)
Chinchillidae	<i>Chinchilla lanigera</i> ¹ (long-tailed chinchilla)

(Continued)

Table 1. (Continued)

Ctenodactylidae	<i>Ctenodactylus gundi</i> (northern gundi)
Ctenomyidae	<i>Ctenomys sociabilis</i> (social tuco-tuco)
Cuniculidae	<i>Cuniculus paca</i> (lowland paca)
Dasyproctidae	<i>Dasyprocta punctata</i> (punctate agouti)
Dinomyidae	<i>Dinomys branickii</i> (pacarana)
Echimyidae	<i>Capromys pilorides</i> (Desmarest's hutia), <i>Myocastor coypus</i> (nutria)
Erethizontidae	<i>Erethizon dorsatum</i> (North American porcupine)
Heterocephalidae	<i>Heterocephalus glaber</i> ¹ (naked mole-rat)
Hystricidae	<i>Hystrix brachyura</i> ² , <i>Hystrix cristata</i> (crested porcupine)
Octodontidae	<i>Octodon degus</i> (degu), <i>Octomys mimax</i> (viscacha rat), <i>Tympanoctomys barrerae</i> (plains viscacha rat)
Petromuridae	<i>Petromus typicus</i> (dassie-rat)
Thryonomyidae	<i>Thryonomys swinderianus</i> (Greater cane rat)

The table lists 118 different species on which genomic data have been assembled and made available. Some of these genomes have been annotated (¹) and genes can be looked up directly (<https://www.ncbi.nlm.nih.gov/gene/>). All of them are accessible using NCBI's blastn suite (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) in cases where genes have not been assigned or more refined individual search options are necessary. The 118 different species are represented by 217 genome assemblies that have been published as of October 2022. For clarity, the species are grouped into their respective families and rodent suborders based on previous suggestions by D'Elia & co-workers (2019). The species cover a total of 29 out of 35 families within the order Rodentia. It should be noted that laboratory strains of *Mus musculus* as well as of *Rattus norvegicus* are not considered to be different species in this overview. Interestingly, despite the genus *Mus*, which is represented by nine different species, several other genera are represented with multiple species as well, such as *Peromyscus* (nine species listed), *Microtus* (eight species), *Acomys* (five species), *Dipodomys* (four species), *Marmota* (five species) and *Cavia* (three species) allowing a side-by-side comparison of genes in closely related genomes. The Latin species name is followed by the English name equivalent if applicable. (²) Indicates cases in which no English name equivalent exists.

using the combination of the search terms 'ACTB and rodentia' or 'SCNN1D and rodentia'). Annotated genomes display the identified gene in its genomic context, whereas the number of potentially available sequence genomes is by far larger and can be accessed using the 'blastn' option (see below).

Despite the use of algorithms that are set up to identify distinct sequence features during the annotation process (for details on the annotation process and used algorithms, such as Splign, Prosplign and Gnomon, see https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/), three problems seem to impede an exhaustive annotation during genome assembly and annotation: (1) a lack of sequence information due to sequencing gaps in the process of genome assembly; (2) an increase in a nucleotide bias for GC, preferably in the third position of synonymous coding triplets (e.g. *NLGN4*, Maxeiner et al., 2020); and (3) the presence of tandem repeats in intronic as well as in intergenic sequences, which likely complicate proper placement of smaller exon-coding DNA sequences surrounded by introns bearing these repeats. The lack of sequence information will be discussed in more depth in the following section. Recently, we published two studies on genes with peculiar evolutionary fates in rodents (*NLGN4* and *SCNN1D*) and both genes share an accumulation of GC nucleotides and repetitive sequences (*NLGN4*, Maxeiner et al., 2020, 2022; *SCNN1D*, Gettings et al., 2021). *NLGN4* is a neural cell adhesion molecule that has suffered sub-

stantial sequence variation during its evolution on a sub-branch of the rodent order, the suborder Supramyomorpha, and the order Lagomorpha (Maxeiner et al., 2020, 2022). The accumulation of GC nucleotides and repetitive sequences in the Supramyomorpha occurs alongside the erosion of the PAR and its gene content which is otherwise well-preserved in mammals. Whether this is a consequence or a cause of the PAR erosion needs to be determined. An increase in overall GC content and the presence of long, repetitive stretches of genomic DNA hampers reliable sequencing, leaving only partial and consequently incomplete gene sequences being placed into genomic assemblies. Accordingly, the recent publication of a full human genome (Nurk et al., 2022) must have come as a surprise given that the human reference genome had already been published two decades ago and suggested that every human gene or other feature (e.g. microRNAs, long non-coding RNAs, etc.) was properly assigned and annotated (Lander et al., 2001). Within these two decades, sequences that were branded 'unsequenceable' due to repetitive stretches or high GC content have become less of a challenge, for example, in RNA-Seq applications, thermostable group II intron reverse transcriptases (TGIRTs) have provided high fidelity and processivity (Belfort & Lambowitz, 2019; Xu et al., 2019). These recent improvements consequently warrant that more recent genome annotations and assemblies might incorporate fewer mistakes and more accurate sequence results. Thus, researchers might opt

for searches within more recent genome assemblies rather than older ones in cases where multiple genome assemblies of the same species are available. If gene search is obstructed by a lack of annotation in any given genome, the information given for closely related species of the same genus or family could be retrieved and submitted to NCBI's blast suite ('Basic Local Alignment Search Tool'; <https://blast.ncbi.nlm.nih.gov/Blast.cgi>) in order to identify the non-annotated gene. For the genus *Mus*, that is, mice in a narrower sense, information from the genomes of *M. pahari* and *M. caroli* (Thybert et al., 2018) or *M. spretus* and *M. castaneus* (Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK) could be used (see also Table 1). An even more refined look into genomic differences of laboratory strains is also available (Lilue et al., 2018), considering strategies of breeding into different backgrounds of laboratory mice. Potential differences between breeding facility-specific mouse strains such as C57BL/6N (NIH, National Institutes of Health, USA) or C57BL/6J (JAX, The Jackson Laboratory, USA) are also considered in physiological studies (Eisfeld et al., 2019; Kendal & Schacht, 2014; Mekada & Yoshiki, 2021).

The quality of the deposition of full genomes is mirroring the technical possibilities at a given time. An eminent example from our own research is the annotation of the guinea pig (*Cavia porcellus*) genome (Cavpor3.0; Broad Institute, Cambridge, MA, USA). Guinea pigs are representatives of rodents outside the above-mentioned suborder Supramyomorpha (D'Elia et al., 2019) and relevant model animals in biomedical research. For example, the guinea pig *NLGN4* gene is highly similar to its human orthologue, whereas *NLGN4* in laboratory mice is not (Bolliger et al., 2008; Jamain et al., 2008; Maxeiner et al., 2020). The *SCNN1D* gene is present and functional in humans and guinea pigs, but not in mice (Gettings et al. 2021). The Cavpor3.0 release dates back to 2008, and sequencing gaps which are present in this genome are yet to be amended. Filling these gaps, however, will likely be up to efforts by individual research labs studying their respective genes of interest (for *SCNN1* genes see Gettings et al., 2021), and often involves a tedious undertaking by PCR cloning and Sanger sequencing. However, to compensate for the absence of *C. porcellus* sequence information (Assembly ID: 304568), genomic data from the Montane guinea pig, *C. tschudii* (Assembly ID: 8252328; Broad Institute, Cambridge, MA, USA), and the Brazilian guinea pig, *C. aperea* (Assembly ID: 1067048; Leibnitz Institute for Zoo and Wildlife Research, Berlin, Germany) are available but not yet annotated. Genomic information is deposited with GenBank and available for sequence searches at <https://www.ncbi.nlm.nih.gov/assembly>.

Problem 2 – Not all gene annotations are necessarily correct

Algorithms that aim to determine the likely start and the end of exons by comparison of splice acceptor and splice donor sites are used in the process of automated gene annotation (e.g. Splign, Kapustin et al., 2008). The presence of exons comprising 5' untranslated sequences are potentially inferred by comparison of already deposited sequences. This annotation process often deems a sequence as 'predicted', which means that it formally lacks experimental validation. Accepting 'predicted' as 'validated' might in many, perhaps even most cases, not affect the outcome of an experimental procedure. In some cases, however, subtle but not scrutinised deviations in a coding sequence can affect the entire outcome of a physiological study. This is the case when 'arbitrary' exons are annotated by algorithms that seek sequence similarity to determine potential exons upstream or downstream of genomic sequence gaps. These exons serve as 'placeholders' within a coding sequence, based on three characteristics: (1) a comparable size to an existing exon in a related species; (2) being framed by appropriate splice acceptor- (simplified: a pyrimidine-rich stretch followed by AG) and donor-like motifs (simplified: GT and in rare cases GC) (Burset et al., 2001); and (3) the absence of a stop codon terminating the anticipated coding region. We recently came across two such cases. *ANOS1* is a human X-chromosomal gene that is implicated in Kallmann syndrome (anosmia and hypogonadotrophic hypogonadism) (De Castro et al., 2014). Generally, its coding region covers 14 exons in over 20 representative primate and rodent species that we have inspected. It is also annotated in the *C. porcellus* genome with a total of 14 exons (Gene ID: 100712973). However, close comparison with the human and other rodent sequences revealed that exon 8 is hidden within an approximately 7 Kb large sequencing gap between exon 7 and exon 9. However, exon 8 has been predicted to be localised downstream of this gap and upstream of the subsequent exon 9 to fit into what is present from the rest of the coding sequence. A more severe case happened for the gene *SCNN1B*, coding for the β -subunit of ENaC, in the coelacanth *Latimeria chalumnae* (Gene ID: 102355590). The coding region of *SCNN1B* covers 12 exons (Gettings et al., 2021, on the numbering of exons). Exons 2 and 13 encode the first and second transmembrane region of the ion channel subunit, respectively. In this example, a major gap (approx. 27 Kb) in the genomic sequence of the coelacanth obscures the proper sequence of exons 3 and 4, which have been replaced in the annotated genomic sequence by a total of eight novel exons, some of which are not even flanked by proper splice sites. This prediction resulted in a presumptive coding sequence

remaining uninterrupted by a stop codon. This example demonstrates that it is crucial to perform due diligence in retrieving and gathering sufficient information about common gene features (e.g. exon sizes, number of exons) and compare them with closely related species in order to identify such annotation errors. The section 'How to avoid nasty surprises' below is accompanied by supporting material that describes a workflow of how to search for and handle sequence information. This workflow uses the sequence information of *SCNN1D* in guinea pig and exemplifies the extent to which knowledge of gene features helps to scrutinise the sequence quality of automated annotation results.

Problem 3 – Transcripts contain automated corrections

In our previous study on the *SCNN1D* gene in rodents (Gettings et al., 2021), we scrutinised the deposited sequences from the 'squirrel-like' suborder Sciuromorpha. *SCNN1D* genes are annotated in the genomes of two marmots, *Marmota flaviventris* (Gene ID: 114102187) and *Marmota monax* (Gene ID: 124079371), suggesting not only the presence of the genes but also, upon translation of the coding sequence, functional *SCNN1D* proteins in marmots (Fig. 2). Closer inspection using sequence comparison as well as a comparison with the general organisation of *SCNN1D* genes in rodents reveals that those sequences violate consistent gene features. The representation of both annotated *SCNN1D* genes of *M. flaviventris* and *M. monax* on their respective NCBI summary page suggests functional genes which are the result of amendments and corrections. Sequence alignments of the underlying genomic sequences with the respective nucleotide sequences representing the assigned exon sequences (*M. flaviventris*, XM_046424888.1; *M. monax*, XM_02794752.1) reveal that the assignment of the coding region consistently violates the integrity of conserved exons, in particular due to the addition of one or two nucleotides (rather than multiples of three) with the aim of generating a proper reading frame 'made to fit'. In addition, the reading frame of both species, which contained a premature stop codon within exon 10, has been corrected (Fig. 2C, right magnification). The corresponding protein sequences contain an 'X', which indicates an unspecified amino acid encoded at this position instead of the premature termination effectively now terminating at the subsequent downstream in-frame stop codon in exon 13 (*M. flaviventris*, XP_027803253.1; *M. monax*, XP_046280844.1). Additionally, the coding sequence of *M. monax* was corrected within exon 5 by adding two unspecified nucleotides to catch up with the reading frame that yields the most likely translational product and resembles a functional gene sequence (Fig. 2C, left magnification).

Given phylogenetic relationships, the presence of mutations, deletions or insertions at the same positions in the genomes of closely related species (e.g. the TAG stop codon in the two marmot species) clearly highlights that these genetic changes accumulated at some point earlier in evolution, were consistently passed on and are not the result of sequencing errors.

Automated amendments and annotation gaps display two facets of a wider problem: (1) genes can be present in a given species despite not being annotated and (2) the presence of a gene can be falsely predicted by corrections 'made to fit' from highly similar sequences, but the gene might actually be 'decaying' and rather representing a pseudogene. It is useful to consult genomic information available from other members of a genus or within a family to prove the validity of potential corrections. For rodents, Table 1 summarises all currently available genomes.

Problem 4 – Mismatches between gene, mRNA and protein sequences

Links between genomic sequences, transcripts and proteins in annotated genomes help to easily switch between each feature, suggesting that accurate translation between genomic sequence, coding sequence and protein sequence is retained. In the case of the guinea pig *SCNN1B* and *SCNN1G* genes, both have been studied using 5'RACE (rapid amplification of cDNA ends; Frohman et al., 1988) to identify potential upstream exons using intron-spanning primers for PCR (Gettings et al., 2021). The results corresponded to the actual genomic sequence deposited to GenBank (*SCNN1B*, Gene ID 100270805, and *SCNN1G*, Gene ID 100270806), whereas confusion arose regarding the translated sequences. In the case of *C. porcellus SCNN1B*, the third, fourth and seventh to last amino acids were mistranslated, in the case of *SCNN1G*, four out of the five C-terminal amino acids differed from the underlying genomic sequence (protein sequences NP_001166534.1 and NP_001166535.1, respectively). Sequence comparison with the genome of *C. tshudii* confirmed the correctness of the underlying *C. porcellus* genomic sequence, thus leaving the reasons for these translational mistakes unexplainable. Consequently, it is worth additionally confirming translated sequences using software tools such as Translate (<https://web.expasy.org/translate>), which translates nucleotide sequences into amino acid sequences. One should also keep in mind that sequence information from cDNA experiments might differ due to mRNA editing. A post-transcriptional modification of adenosine to inosine has been reported, for example, in a number of neurotransmitter receptors and ion channels of the nervous system (Hood & Emeson, 2012). Many more

of these modifications have been identified and these studies are an effort to understand the epitranscriptome of normal or diseased cells (for review see: Arzumian et al., 2022). To assess changes between nucleotide sequences deriving from genomic sequences and cDNA sequences, a comparison between closely related species might reveal whether these changes are consistent or not.

Problem 5 – Splicing patterns

Researchers are often confronted with the possibility of alternative splicing, leaving them indecisive as to which of the potentially numerous ‘predicted’ transcripts to choose from. Alternative splicing emerges in different forms, such as exon skipping, intron inclusion, mutually exclusive exon usage and changes to the splicing

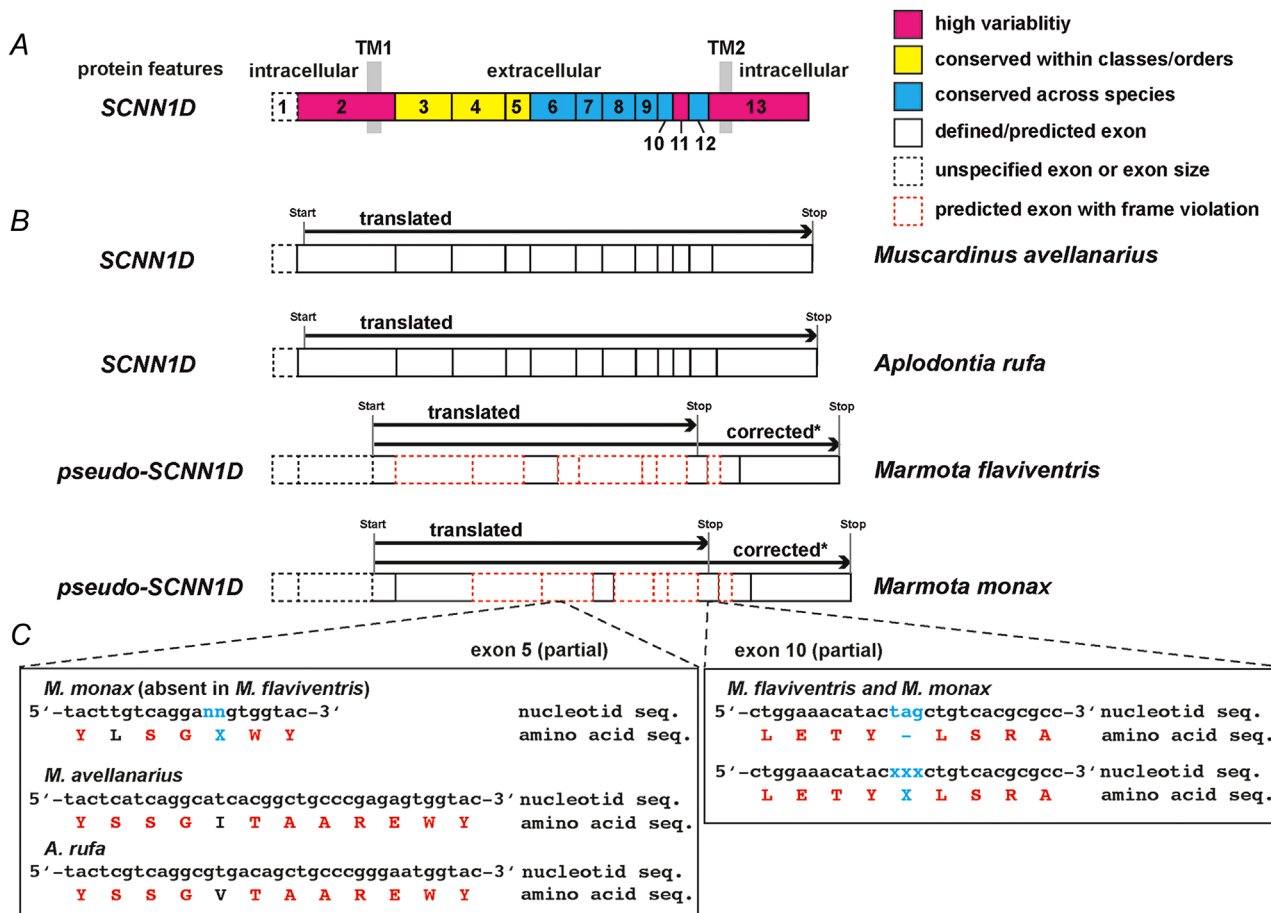


Figure 2. Misinterpretation of pseudogenes as genes during the annotation process
 The process of misinterpreting pseudogenes as genes is exemplified for two cases of the *SCNN1D* gene in *Marmota flaviventris* and *Marmota monax*. **A**, a general representation of the *SCNN1D* gene based on Gettings et al. (2021). Topological features are assigned to distinct exons of the coding region, for example, both transmembrane regions (TM1/TM2) to exons 2 and 13, respectively. For some species, exons upstream of exon 2 have been identified by coding for the translational start, which are represented by boxes with a dashed border. The size of each exon varies across different classes/orders, but the sizes of exons 6 to 10 as well as exon 12 remain generally highly conserved (for details see Gettings et al., 2021). **B**, within the ‘squirrel-like’ suborder of rodents, the Sciuromorpha, the family of marmots has lost a functional *SCNN1D* gene, whereas the related species *Apodontia rufa* and *Muscardinus avellanarius*, both from a different evolutionary branch within this group, share all gene features allowing them to likely express a functional *SCNN1D* protein. These features include: a single reading frame starting in exon 2 to a translational stop in exon 13; exon boundaries identical to the very conserved regions, and/or differences not violating the reading frame; and furthermore, the consistent presence of the same features in other species closely related to them. **C**, the box on the left side is a partial depiction of the nucleotide sequence of exon 5 and its translation into the amino acid sequence of *M. monax*, *M. avellanarius* and *A. rufa*. The box on the right side shows parts of the nucleotide and corresponding amino acid sequence of exon 10 from *M. flaviventris* and *M. monax*.

of 5' and 3-UTR regions. Many receptors or channels are subject to alternative splicing, such as voltage-gated calcium channels (Lipscombe et al., 2013), transient receptor potential channels (Vázquez & Valverde, 2006), acid-sensing ion channels (Babini et al., 2002) or glycine receptors (Lemmens et al., 2022), just to name a few. If one is studying a gene, which is known from published work based on the corresponding human or mouse genes to retain a distinct splicing pattern, one should consider these splice versions before taking into account other predicted transcripts. If a plethora of alternatively spliced transcripts is displayed listing transcripts with potentially skipped exons, additionally included (novel) exons or a mixture of both, one should consider analysing the impact of the absence or presence of these exons on the reading frame, possibly leading to a preliminary translational stop. One is also advised to check the literature on specific splicing products bearing one question in mind: Have all splice products previously been validated?

For instance, in the gene family of latrophilins (adhesion G protein-coupled receptors, *ADGRL1-3*), the gene for latrophilin-2 (*ADGRL2*) displays 50 alternatively spliced exons which are classified as 'predicted' in mouse, 27 in guinea pig and only seven in the American beaver (e.g. Gene ID: 99633 for mouse *Adgrl2*; guinea pig, 100713240; American beaver, 109686631). So far, however, the inclusion or absence of only one mini-exon (15 bases) has been quantitatively assessed and confirmed by PCR (Boucard et al., 2014); others have been tested functionally (Li et al., 2020; Ovando-Zambrano et al., 2019). The combinatorics of different exon combinations as a result of alternative splicing demands *in vitro* or *in vivo* confirmation. If alternative splicing is of major concern for a given project, previous studies might be available that address this issue, helping to focus on those transcripts which are more likely expressed in the tissues or cell types that one aims to study. A caveat, however, is to entirely rely on RNA-Seq data, because these solely match results to reference transcriptomes of curated databases and might ignore the presence of potential non-referenced transcripts (Morillon & Gautheret, 2019).

How to avoid nasty surprises

In the following section, we aim to help jumpstart any kind of project that revolves around a protein of interest of which its cDNA has not already been cloned and therefore curated. The following information and resources reflect the strategies that we have been employing to study individual genes. It should be noted that the use of web-based tools/algorithms from the websites mentioned below are, in this regard, biased. A plethora of additional websites are available that fulfil similar purposes and can be employed. In addition to these strategies, one should

keep in mind that some labs might have already made the effort to study the same protein of interest and are often happy to share their insights and expertise.

As PubMed searches (<https://pubmed.ncbi.nlm.nih.gov/>) have established themselves as the eminent strategy to dive into the literature on a particular topic, a gene or protein of interest, the NIH also hosts one of the biggest genetic sequence databases: GenBank. GenBank is an annotated collection of all publicly available DNA sequences (Benson et al., 2012; <https://www.ncbi.nlm.nih.gov/genbank/>).

To retrieve sequence information on a particular gene, one should search under the tab 'Gene' for the gene of interest in combination with the name of the species. A summary of such a workflow is depicted in Fig. 3. A more detailed step-by-step presentation studying the *SCNN1D* gene in guinea pig can be found in the Supporting information. This workflow contains not only the search for the *SCNN1D* sequence but also strategies of how to handle the retrieved information.

Any search will lead the researcher eventually to lists of potential nucleotide sequences. Please note that most of the annotations regarding sizes of the first and last

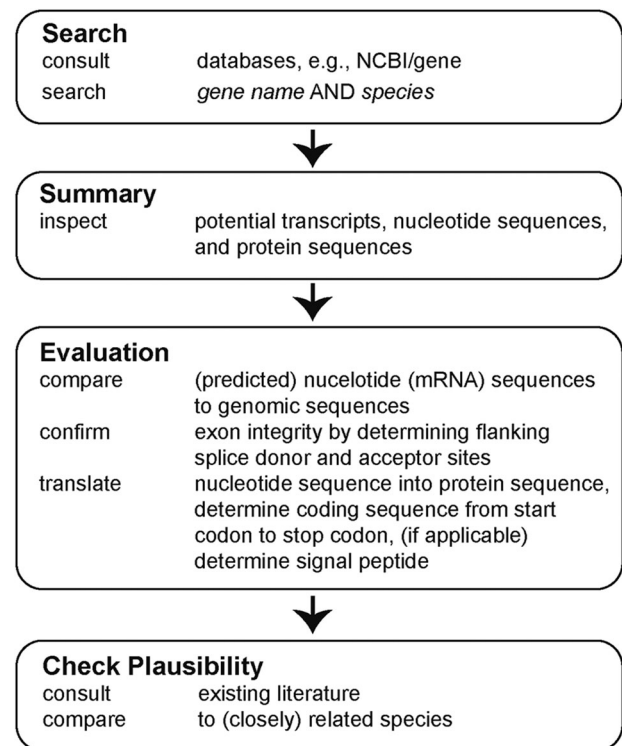


Figure 3. Flow chart depicting a process for retrieving gene sequence information

This flow chart summarises four essential steps to identify gene sequence information. While the boxes 'Search' and 'Summary' help to identify a gene of interest, suggestions in the boxes 'Evaluation' and 'Check Plausibility' help to scrutinise the retrieved data and sequence information.

exons as well as potential cases of alternative splicing are the result of algorithms and deem the status of the corresponding nucleotide sequences as 'predicted'. These sequences have not been experimentally validated. In contrast, lists of, for example, mouse, rat and human nucleotide sequences might display older entries from individual research labs that had validated, for instance, transcriptional start sites by 5'RACE experiments, or determined potential constitutive and/or alternative splice sites experimentally. If physiological changes in protein products resulting from alternative splicing from the very same gene are anticipated or have been reported previously, the underlying nucleotide sequences to be used should be considered with caution.

Within the last decade, synthetic biology has been thriving and custom sequences can be purchased ready to use for a comparably low price. Hence, as a prerequisite to any application one should properly inspect, align and compare the sequences retrieved from databases with what has been available and successfully used in experimental studies to avoid surprises such as edited/redacted sequences which we have alluded to in the section on problem 4.

The determination of splicing patterns and framing individual exons can be easily accomplished using simple sequence comparison tools, such as MultAlin (Corpert, 1988) or Clustal Omega on the EMBL-EBI website (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) (Sievers & Higgins, 2014), with default settings for either protein or nucleotide (DNA/RNA) sequence alignment. For a successful assessment of the splicing pattern, a basic knowledge of sequence motifs leading up to an exon, splice acceptor sites, or following a proper exon, splice donor sites, is of the utmost importance. As a rule of thumb, splice acceptor sites are pyrimidine-rich stretches of genomic DNA followed by 'AG', and for the splice donor, the exon is flanked by 'GT', in most instances, leading to the 'GT-AG' rule as starting and ending sequences of introns (Hastings & Krainer, 2001). There might be rare exceptions but the only notable one is the observation of a 'GC' instead of a 'GT' which accounts for approximately 1 in 120 splice sites in human and mouse genes but is also found in *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana* (Sheth et al., 2006). Should the coding sequence and the splicing pattern already be known, for example, from a related species, one can also align this sequence to the genomic or nucleotide sequence of the species of interest (see Table 1 for potential genomes in rodents). This approach emerges as quite powerful in cases where dramatic changes have occurred to the overall genomic region in which the gene is supposed to be located. It was this very approach that recently led us to identify a fusion of two exons to a super-exon in the guinea pig *SCNN1D* gene (Gettings et al., 2021). Puzzled by this observation, we compared

closely related species with the guinea pig within the 'porcupine-like' rodent suborder Hystricomorpha and were able to confirm this genomic change as valid due to the consistent loss of functional splice donor and acceptor sites between these two exons.

Once an assignment to individual exons has been accomplished, a translation of the anticipated coding region into the correct amino acid sequence is necessary to confirm or reject the hypothetical coding region, that is, is the full reading frame intact or does it result in a truncated protein due to a preliminary stop codon? Web-based tools such as Translate (<https://web.expasy.org/translate/>) are available and will provide all three sense and antisense reading frames. Although a determination of the translational stop is simple, that is, the translational product will inevitably terminate at some point, the determination of the translational start, however, might not be that obvious. All frames will display a presumptive methionine start codon and display the longest possible reading frame. This does not mean that this very amino acid is, indeed, the translational start; another methionine downstream within the sequence could mark the potential N-terminus of the protein. Alternatively, some frames do not display any methionine to start from, suggesting that the beginning of the very first exon still lacks proper annotation, or that one or more upstream exons bearing the translational start have not yet been identified.

Physiologists focusing on ion channels or transporters are often confronted with the cloning of a particular transmembrane protein. These require a signal peptide so that the nascent protein sequence is inserted properly into the lipid bilayer. Programs such as SignalP-5.0 (Almagro Almenteros et al., 2019; <https://services.healthtech.dtu.dk/service.php?SignalP-5.0>) determine the potential cleavage site using algorithms and statistics, separating the signal peptide from the rest of the protein. It might add a layer of confidence when the translational product is further analysed in that regard.

In cases where proper exon assignment has been performed and a considerable degree of sequence similarity has been determined between the newly retrieved and the already curated sequences, full-length translational products might not solely be present in a single reading frame. In this case, the second and third reading frames should be inspected for sequence homology; for example, for the presence of conserved C-terminal motifs if applicable.

Changes of the reading frame are often the result of misinterpreted splice sites or could be attributed to sequencing mistakes in regions of highly similar nucleotides that have been misinterpreted during the sequencing process. This, for instance, is the case for the currently deposited genomic sequence for *SCNN1D* in *C. porcellus* (guinea pig, Gene ID: 100714892). Our own

sequencing efforts resulted in a sequence with a different C-terminus (Nucleotide Acc. No.: MN187539; Gettings et al., 2021) that corresponds to the more recently released but not yet annotated sequence of the Montane guinea pig *C. tschudii* (Nucleotide Acc. No.: PVKK010006848.1). This could possibly be explained by the different release dates of both genomes employing different sequencing approaches. The genome of *C. porcellus* was made available in 2008 using Sanger sequencing and that of *C. tschudii* in 2019 using IlluminaSeq. To reconcile for such striking differences a sequence alignment of closely related species, such as found in Table 1, may therefore be supportive and beneficial in identifying the most likely candidate sequence to any project.

Conclusions

In 1929, Danish Physiologist and Nobel Laureate August Krogh stated: 'For a large number of problems there will be some animal of choice or a few such animals on which it can be most conveniently studied.' (Krogh, 1929). Almost one century later, advances in genome sequencing techniques and rapidly growing genomic data from many different species pave the way for comparative physiological studies addressing a large number of questions and problems. However, the ever-growing amounts of genomic data in public resources warrant careful curation and inspection. Gaps and errors in gene annotation, automated sequence corrections, predicted splicing patterns and mismatches between genes, transcripts and proteins, should be carefully considered prior to functional physiological studies which rely on such sequence information.

References

- Akane, A., Shiono, H., Matsubara, K., Nakahori, Y., Seki, S., Nagafuchi, S., Yamada, M., & Nakagome, Y. (1991). Sex identification of forensic specimens by polymerase chain reaction (PCR): Two alternative methods. *Forensic Science International*, **49**(1), 81–88.
- Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, **37**(4), 420–423.
- Arzumanian, V. A., Dolgalev, G. V., Kurbatov, I. Y., Kiseleva, O. I., & Poverennaya, E. V. (2022). Epitranscriptome: Review of Top 25 most-studied RNA modifications. *International Journal of Molecular Sciences*, **23**(22), 13851.
- Babini, E., Paukert, M., Geisler, H. S., & Grunder, S. (2002). Alternative splicing and interaction with di- and polyvalent cations control the dynamic range of acid-sensing ion channel 1 (ASIC1). *Journal of Biological Chemistry*, **277**(44), 41597–41603.
- Bhatt, M., Di Iacovo, A., Romanazzi, T., Roseti, C., Cinquetti, R., & Bossi, E. (2022). The “www” of *Xenopus laevis* oocytes: The why, when, what of *Xenopus laevis* oocytes in membrane transporters research. *Membranes*, **12**(10), 927.
- Belfort, M., & Lambowitz, A. M. (2019). Group II Intron RNPs and reverse transcriptases: From retroelements to research tools. *Cold Spring Harbor perspectives in biology*, **11**(4), a032375.
- Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, **40**(D1), D48–D53.
- Bolliger, M. F., Pei, J., Maxeiner, S., Boucard, A. A., Grishin, N. V., & Südhof, T. C. (2008). Unusually rapid evolution of Neuroligin-4 in mice. *PNAS*, **105**(17), 6421–6426.
- Boucard, A. A., Maxeiner, S., & Südhof, T. C. (2014). Latrophilins function as heterophilic cell-adhesion molecules by binding to teneurins: Regulation by alternative splicing. *Journal of Biological Chemistry*, **289**(1), 387–402.
- Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2001). SpliceDB: Database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Research*, **29**(1), 255–259.
- Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research*, **16**(22), 10881–10890.
- De Castro, F., Esteban, P. F., Bribián, A., Murcia-Belmonte, V., García-González, D., & Clemente, D. (2014). The adhesion molecule anosmin-1 in neurology: Kallmann syndrome and beyond. *Advances in Neurobiology*, **8**, 273–292.
- D'Elia, G., Fabre, P.-H., & Lessa, E. P. (2019). Rodent systematics in an age of discovery: Recent advances and prospects. *Journal of Mammalogy*, **100**(3), 852–871.
- Eisfeld, A. J., Gasper, D. J., Suresh, M., & Kawaoka, Y. (2019). C57BL/6J and C57BL/6NJ mice are differentially susceptible to inflammation-associated disease caused by influenza A virus. *Frontiers in Microbiology*, **9**, 3307.
- Frohman, M. A., Dush, M. K., & Martin, G. R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences*, **85**(23), 8998–9002.
- Fronius, M., Bogdan, R., Althaus, M., Morty, R. E., & Clauss, W. G. (2010). Epithelial Na⁺ channels derived from human lung are activated by shear force. *Respiratory Physiology & Neurobiology*, **170**(1), 113–119.
- Gettings, S. M., Maxeiner, S., Tzika, M., Cobain, M. R. D., Ruf, I., Benseler, F., Brose, N., Krasteva-Christ, G., Vande Velde, G., Schönberger, M., & Althaus, M. (2021). Two functional epithelial sodium channel isoforms are present in rodents despite pronounced evolutionary pseudogenization and exon fusion. *Molecular Biology and Evolution*, **38**(12), 5704–5725.
- Hastings, M. L., & Krainer, A. R. (2001). Pre-mRNA splicing in the new millennium. *Current Opinion in Cell Biology*, **13**(3), 302–309.
- Hood, J. L., & Emeson, R. B. (2012). Editing of neurotransmitter receptor and ion channel RNAs in the nervous system. *Current Topics in Microbiology and Immunology*, **353**, 61–90.

- Jamain, S., Radyushkin, K., Hammerschmidt, K., Granon, S., Boretius, S., Varoquaux, F., Ramanantsoa, N., Gallego, J., Ronnenberg, A., Winter, D., Frahm, J., Fischer, J., Bourgeron, T., Ehrenreich, H., & Brose, N. (2008). Reduced social interaction and ultrasonic communication in a mouse model of monogenic heritable autism. *PNAS*, **105**(5), 1710–1715.
- Kapustin, Y., souvorov, A., Tatusova, T., & Lipman, D. (2008). Splign: Algorithms for computing spliced alignments with identification of paralogs. *Biology Direct*, **3**(1), 20.
- Kendall, A., & Schacht, J. (2014). Disparities in auditory physiology and pathology between C57BL/6J and C57BL/6N substrains. *Hearing Research*, **318**, 18–22.
- Krogh, A. (1929). The progress of physiology. *Science*, **70**(1809), 200–204.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczkzy, J., ... LeVine, R. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.
- Lemmens, V., Thevelein, B., Vella, Y., Kankowski, S., Leonhard, J., Mizuno, H., Rocha, S., Brône, B., Meier, J. C., & Hendrix, J. (2022). Hetero-pentamerization determines mobility and conductance of Glycine receptor $\alpha 3$ splice variants. *Cellular and Molecular Life Sciences*, **79**(11), 540.
- Lilue, J., Doran, A. G., Fiddes, I. T., Abrudan, M., Armstrong, J., Bennett, R., Chow, W., Collins, J., Collins, S., Czechanski, A., Danecek, P., Diekhans, M., Dolle, D. D., Dunn, M., Durbin, R., Earl, D., Ferguson-Smith, A., Flicek, P., Flint, J., ... Keane, T. M. (2018). Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nature Genetics*, **50**(11), 1574–1583.
- Li, J., Xie, Y., Cornelius, S., Jiang, X., Sando, R., Kordon, S. P., Pan, M., Leon, K., Südhof, T. C., Zhao, M., & Araç, D. (2020). Alternative splicing controls teneurin-latrophilin interaction and synapse specificity by a shape-shifting mechanism. *Nature Communications*, **11**(1), 2140.
- Lipscombe, D., Andrade, A., & Allen, S. E. (2013). Alternative splicing: Functional diversity among voltage-gated calcium channels and behavioral consequences. *Biochimica Et Biophysica Acta*, **1828**(7), 1522–1529.
- Maxeiner, S., Sester, M., & Krasteva-Christ, G. (2019). Novel human sex-typing strategies based on the autism candidate gene NLGN4X and its male-specific gametologue NLGN4Y. *Biol Sex Differ*, **10**(1), 62.
- Maxeiner, S., Benseler, F., Krasteva-Christ, G., Brose, N., & Südhof, T. C. (2020). Evolution of the autism-associated neuroligin-4 gene reveals broad erosion of pseudo-autosomal regions in rodents. *Molecular Biology and Evolution*, **37**(5), 1243–1258.
- Maxeiner, S., Gebhardt, S., Schweizer, F., Venghaus, A. E., & Krasteva-Christ, G. (2021). Of mice and men – and guinea pigs? *Annals of Anatomy*, **238**, 151765.
- Maxeiner, S., Benseler, F., Brose, N., & Krasteva-Christ, G. (2022). Of humans and gerbils- independent diversification of neuroligin-4 into X- and Y-specific genes in primates and rodents. *Frontiers in Molecular Neuroscience*, **15**, 838262.
- Mekada, K., & Yoshiki, A. (2021). Substrains matter in phenotyping of C57BL/6 mice. *Experimental Animals*, **70**(2), 145–160.
- Morillon, A., & Gautheret, D. (2019). Bridging the gap between reference and real transcriptomes. *Genome Biology*, **20**(1), 112.
- Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bizkadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., ... Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, **376**(6588), 44–53.
- Ooi, A., Wong, A., Esau, L., Lemtiri-Chlieh, F., & Gehring, C. (2016). A Guide to Transient Expression of Membrane Proteins in HEK-293 Cells for Functional Characterization. *Frontiers in Physiology*, **7**, 300.
- Ovando-Zambrano, J. C., Arias-Montaño, J. A., & Boucard, A. A. (2019). Alternative splicing event modifying ADGRL1/latrophilin-1 cytoplasmic tail promotes both opposing and dual cAMP signaling pathways. *Annals of the New York Academy of Sciences*, **1456**(1), 168–185.
- Papke, R. L., & Smith-Maxwell, C. (2009). High throughput electrophysiology with *Xenopus* oocytes. *Combinatorial Chemistry & High Throughput Screening*, **12**, 38–50.
- Sievers, F., & Higgins, D. G. (2014). Clustal omega. *Current Protocols in Bioinformatics*, **48**(1), 3.13.1–16.
- Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., & Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research*, **34**(14), 3955–3967.
- Thybert, D., Roller, M., Navarro, F. C. P., Fiddes, I., Streeter, I., Feig, C., Martin-Galvez, D., Kolmogorov, M., Janoušek, V., Akanni, W., Aken, B., Aldridge, S., Chakrapani, V., Chow, W., Clarke, L., Cummins, C., Doran, A., Dunn, M., Goodstadt, L., ... Flicek, P. (2018). Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Research*, **28**(4), 448–459.
- Vázquez, E., & Valverde, M. A. (2006). A review of TRP channels splicing. *Seminars in Cell & Developmental Biology*, **17**(6), 607–617.
- Wichmann, L., Dulai, J. S., Marles-Wright, J., Maxeiner, S., Szczesniak, P. P., Manzini, I., & Althaus, M. (2019). An extracellular acidic cleft confers profound H⁺-sensitivity to epithelial sodium channels containing the δ -subunit in *Xenopus laevis*. *Journal of Biological Chemistry*, **294**(33), 12507–12520.
- Xu, H., Yao, J., Wu, D. C., & Lambowitz, A. M. (2019). Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Scientific Reports*, **9**(1), 7953.
- Zaffalon, S., Latz, A., Krasteva-Christ, G., & Maxeiner, S. (2019). Sex identification in horses (*Equus caballus*) based on the gene pair NLGN4X/NLGN4Y. *Animal Genetics*, **50**(5), 551.

Additional information

Competing interests

The authors have no competing interests to declare.

Author contributions

S.M. and M.A. wrote the article. S.M. and M.A. prepared the figures. S.M., G.K.C. and M.A. have edited the article. All authors approved the final version of the manuscript to be published.

Funding

S.M. received funding from the Medical School of Saarland University: HOMFOR2017-19. G.K.C. is supported by the German Research Foundation (DFG): KR4338/1-2 and SFB TRR152 P22. M.A. receives funding from the Ministry of Culture and Science of the State of North Rhine-Westphalia: FKZ 005-2101-0144 and FKZ 005-2211-0043, and DFG: INST 19446/3-1.

Acknowledgements

The authors would like to thank Monika Hollenhorst and Saskia Evers for critical reading of the manuscript. The graphical abstract was created with biorender.com.

Open access funding enabled and organized by Projekt DEAL.

Keywords

gene expression, genomic data, rodents, sequencing

Supporting information

Additional supporting information can be found online in the Supporting Information section at the end of the HTML view of the article. Supporting information files available:

Peer Review History

Supporting Material