Coding variants in mouse and rat model organisms: mousepost & ratpost

Steven Timmermans[1,2,*], Claude Libert[1,2,*]

[1]VIB-Ugent center of inflammation research, Ghent, Belgium

[2]Department of Biomedical Molecular Biology, Ghent university, Ghent, Belgium

[*]corresponding author

Keywords: rat|mouse|genetics|inbred strains|coding|polymorphisms

Abstract

Mice and rats are the most commonly used vertebrate model organisms in biomedical research. The availability of a reference genome in both animals combined with the deep sequencing of several doze of popular inbred lines also provides rich sequence variation data in these species. In some cases, such sequence variants can be linked directly to a distinctive phenotype. In previous work, we created the mouse and rat online searchable databases (“Mousepost” and “Ratpost”) where small variant information for protein coding transcripts in mouse and rat inbred strains can be easily retrieved at the amino acid level. These tools are directly useful in forward genetics strategies or as a repository of existing sequence variations. Here, we perform a comparison between the ”Mousepost” and ”Ratpost” databases and we couple these two tools to a database of human sequence variants ClinVar. We investigated the level of redundancy and complementarity of known variants in protein coding transcripts and found that the large majority of variants is species specific. However, a small set of positions is conserved in an inbred line between both species. We conclude that both databases are highly complementary, but this may change with further sequencing efforts in both species.

Declarations

Funding

Research was funded by the Research Council of Ghent University (GOA program), the Research Foundation Flanders (FWO Vlaanderen) and Flanders Institute for Biotechnology (VIB).

Conflict of interest

On behalf of both authors, the corresponding author states that there is no conflict of interest.

Ethics approval

Not applicable

Consent for Publication

Not applicable

Data availability

All data used is publicly available at the rat genome database, the mouse genomes project, Ensembl, ratpost.be and mousepost.be.

Author contributions

S.T. performed all analyses and wrote the first draft of the manuscript. C.L. guided the research and edited the manuscript.

Introduction

The most popular mammalian model organisms currently in use are the mouse (*Mus musculus*, NCBI Taxon ID 10092) and the rat (*Rattus norvegicus*, NCBI Taxon ID 10116), accounting for over 73% of all vertebrate animals used in research (61% and 12%, respectively) in 2017 in Europe [1]. Both have a rich history in the research community. The rat was the first animal domesticated for research purposes and has been used since 1856 [2, 3], while mice followed in the early 20th century [4]. Both species have their own set of (dis)advantages: mice are small, easy to breed and cost effective to keep while the larger size of rats allows more flexibility in assays to be performed (e.g. multiple samples, higher resolution imaging) [5]. Rats are more often used than mice in cardiovascular research and neurobiology [6]. The genome of both species has been sequenced, mice in 2002 (C57BL/6J as reference strain) [7] and rats in 2004 (BN/SsNHsd as reference strain) [8], as part of a large scale effort to sequence the human genome and the genomes of important model organisms. The publication of these genome sequences has marked the beginning of a new era of research into the genetics and genomics of these and other species, functional studies, evolution, comparison and much more.

About a decade after the first version of each reference genome was published, the widespread adoption and advancement of novel massively parallel sequencing technologies (e.g. using the illumina sequencing platforms) has made genome sequencing faster and less expensive an allowed for the sequencing of genomes of multiple inbred strains of mice and rats. The discovery of sequence variations on a genome wide level and investigation into their effects has hence become possible. The *mouse genomes project* (MGP), launched and performed by the Sanger Institute, sequenced the most used and popular mouse strains, releasing a first set of strain specific variants in 2011 [9]. Currently their dataset includes SNPs and small indels from 36 inbred mouse lines, but the MGP performs additional sequencing on known lines and additional lines will be included, based on data listed on their ftp-server, but they have at the time of writing this work not yet been published. In contrast to the mouse scientific research, which is performed almost exclusively in inbred lines, the rat community

also makes use of outbred stocks, which have an undefined genetic background [10]. The interest in inbred line characterization for rats was therefore traditionally lower and was performed in two main publications, one rat, by Atanur *et al*. [11] including 28 inbred rat strains and a more expanded analysis on 40 lines (including the previous 28) by Hermsen *et al*. in 2015 [12] and is made available through the rat genome database [13]. All information from small variants in mice and rats can be accessed and downloaded at their respective databases (MGP and rat genome database) in variant call format, annotated with the Ensembl variant effect predictor (VEP) and directly accessed through web-based variant browsers.

As information that is specific to the coding sequence and changes to the amino acid sequences cannot be derived immediately or trivially derived for the variant data, our lab has previously processed coding variants, on a per codon basis, in mouse and rat and made these available alongside PROVEAN based predictions on the functional impact on the variant protein, as the “Mousepost” [14] and “Ratpost” [15] databases respectively. These online-available databases offer complementary protein level information to the existing nucleotide level resources available at the mouse genomes project and the rat genome database.

In this study, we compare the data that is present in the “Mousepost” and ”Ratpost” databases. We investigate if the variants in mouse are also found in rat and vice-versa, and if inbred lines from these species can be used interchangeably to some extent or if the variants found are species-specific and/or if the databases provide two complementary datasets that can be used in research.

Results & discussion

Linking mouse and rat variant transcripts

We used the mouse-rat ortholog gene information from the Ensembl website to obtain orthologous gene peptide IDs which were mapped back to related transcript ID. In case of multiple options, only the best match was retained. We filtered both “Mousepost” and “Ratpost” databases on the occurrence of orthologous pairs and we identified 8,154 orthologous pairs that had at least one non-synonymous variant in both rat and mouse protein sequences, out of 18,788 total pairs. However, the

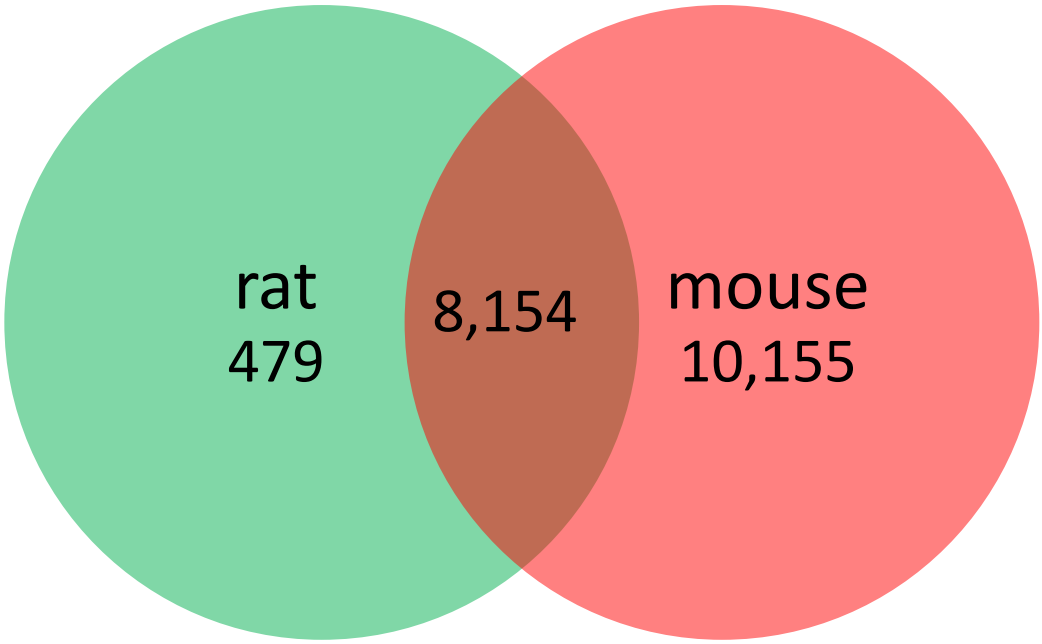

Figure 1 Overlap between mouse and rat orthologous protein coding transcript sequences that have at least one or more sequence variations in at least one rat and one mouse inbred strain. The mouse data has more variants called than the rat. Almost all rat sequences (8154) with a variant in one inbred line have an ortholog in mouse that is also mutated in at least one inbred strain

mouse dataset contains several *recently* wild derived inbred lines. Thise are highly divergent from the C57BL/6J reference strain, which has been bred in captivity for about a century. This level of variation can possibly skew the results due to the fact that these contain a hight amount of private (strain specific) variants. The use of transcripts instead of the actual variants partially compensates for this, but in order to minimize the effect we perform this comparison with the 4 most divergent wild derived lines removed (SPRET/EiJ, PWK/Phj, CAST/EiJ and MOLF/EiJ), which each contain thousands more variants than any of the other strains. This lowers the overlap only slightly, from 8154 to 7659

transcript. As this excludes an overly large bias introduced by wild derived strains, we will include them in our analyses. The total number of variant proteins in rat is almost a subset of the mouse set. A total of 18,788 mouse coding transcripts in the orthologous set has at least one variant, but for rat there are only 8,633 sequences identified in the orthologous set that are deviant from the reference sequence. Thus 94.5% of all rat sequences are shared with mouse, however this does not mean that these sequences are mutated in a similar manner in mouse and rat, only that they are deviant from the reference sequence in both species. It is also interesting to place this into perspective of the total amount of rat variants.

In our previous ”Ratpost” publication we described a total of 12,172 protein coding transcripts with at least one non-synonymous variant [15]. Taking into account that 8,154 transcripts were found to have a variant in at least one inbred line of both strains, and that 479 transcripts only have variants in rat, this means that 3,539 rat transcripts have no assigned ortholog in mouse, either because there is none known, or because it is part of a many-to-many orthology relationship, where one rat gene is orthologous to multiple mouse genes and vice versa. Upon further investigation there were no large pathways or groups in this set, but several genes were found to belong to taste and smell receptors (*Olfr* and *Vmn1r* families).

The vomeronasal, olfactory and also the MHC gene families are known for presence-absence polymorphism and lower than normal reliability in variant calling [17]. In this comparison, as well as in Mousepost an Ratpost, only members that are present in a strain are included, in case of a gene deletion in a strain, the locus is not further investigated, in effect this is the same approach as for a locus that contains no missense or nonsense variants. As we seek to provide an overview that is as complete as possible, these gene families were not excluded, but users should be careful with data from these gene families.

The large difference between mouse and rat in number of variants and variant-containing sequences can likely be attributed to the sequencing efforts performed, which is directly related to their respective use in the research community. Mouse - specific sequencing efforts are headed by a large

institute (the Sanger Institute), and additional sequencing/resequencing is being done giving higher coverage to perform variant calling. By contrast in rat there has been only limited strain-specific sequencing efforts so far and should more sequencing data become available it is likely that the amounts of variants, and transcripts with variants will increase. In addition, only the most used strains have been sequenced, which is only a fraction of the total available inbred strains (hundreds in both rat and mouse [10, 16]), and this number may change if and when more data from other strain becomes available.

Overlap between variant classes for orthologous sequences

When taking into account the different variant classes resulting from small indels and SNPs that were defined in the ”Mousepost” and ”Ratpost” databases, namely stop-gain (SG), stop-loss (SL), and non-stop related mutation (MUT), large differences between mouse and rat variants become readily apparent. Especially for the transcripts that are SG or SL, the overlap between rat and mouse variants is very limited at the classification level, with only 42 transcripts found that result in a nonsense mutation in both species from the 684 rat (6.1%) and 374 mouse (11.2%) transcripts (Table 1) in that category. The same observation can be made for SL: only 8 transcripts are annotated as SL in at least one strain of both species, which is 11.1% for rat and a poor 0.65% for mouse (Table 1). This shows that there is only (very) limited overlap between the variant classes between mouse and rat. It will be difficult to find conditions where there will be genetic equivalence between a mouse and a rat strain for a specific transcript/gene. However, this also allows for a large natural repository with an overview of multiple variants all resulting in loss of function to some degree. We make a detailed comparison of the transcripts that contain an snp/indel in both mouse and rat inbred strains and compare the variants they contain, with a focus on those that make up a missense event (MUT) in at least one strain of the one or both species.

Table 1 number of transcripts assigned in each variant class in mouse and rat. The 8154 transcripts from the orthologous set are compared here, mouse numbers are in the columns and rat data is in the rows of the table.

rat \ mouse	SG	SL	MUT	Total
SG	42	154	488	*684*
SL	5	8	59	*72*
MUT	327	1069	6002	*7398*
Total	*374*	*1231*	*6549*	*8154*

Comparing variant positions

For transcripts that have *conserved* SG or SL variants between species, comparing the size of the truncation or sequence gained is straightforward and was performed in an indirect manner by means of protein length ratios obtained by comparing the strain specific sequence to the respective reference sequence. This compensates directly for variations in length caused by the evolutionary difference between species.

In contrast, comparing positions and substitutions caused by missense mutations (class MUT) was non-trivial due to the differences caused by evolution since speciation. In order to perform comparisons, the reference sequences of each MUT protein were pairwise aligned using Biopython to obtain the optimal global alignment for each orthologous pair. Results from this alignment step were used to create a one-to-one map of mouse vs rat protein positions. Variants were queried from the ”mousepost” and ”ratpost” databases and compared using the positional map that was previously constructed on a per transcript basis. For the MUT transcripts there were a total of 74,964 variant

positions present. Only 619 of these were found in both mouse and rat and 14,599 were found only in the rat and 59,746 were found in the mouse and did not have a rat equivalent (Figure 2).

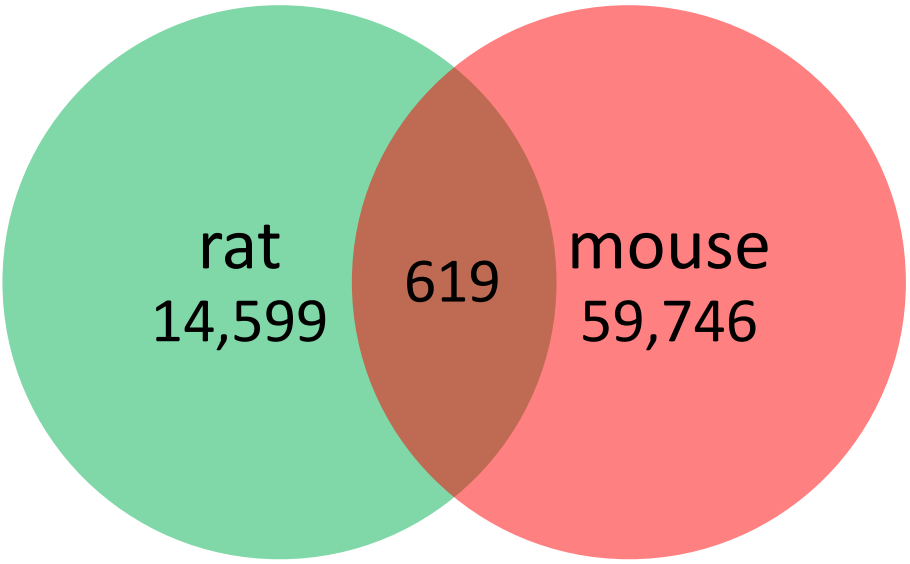

Figure 2. Variants found in mouse-rat orthologs. Only 619 variants occur at equivalent positions in rat and mouse, the large majority of variants in inbred lines is species specific.

A large portion of these variants may result in loss of function (LOF) mutations. Overall, LOF variants can have two main origins, they can be present from natural genetic divergence between species, but can also be the result of relaxed evolutionary pressure on genes that are less important in captivity. The latter group will occur in all species kept in captivity, and thus should be found enriched in the mouse-rat overlap.

These data suggest that the mouse and rat variant databases contain a large amount of complementary information. This increases the chance that at least one of the two model organisms will have a natural variant that matches a human variant. The ”Ratpost” and ”Mousepost” databases can be used to find such variants for human pathological mutation, such as these described in the Clinvar database [18], on the condition that they are protein coding. Furthermore, it also indicates that some studies must be done, or are much easier to do in a species-specific manner that leverages existing variation between strains. An example of this is the LPS resistance observed in the C3H/HeJ mouse strain, which was found to be caused by a mutation in the *Tlr4* gene [19] and can also be found in “Mousepost”. Since this gene is not found to have a protein level mutation in any of the rat strains,

it would not have been possible to find association between *Tlr4* and LPS resistance using the rat model organism without performing active (random) mutagenesis.

Non position equivalent variants

The large majority of mouse and rat variant transcripts do not share a position specific variation. This does not mean that the protein suffers from loss of function in the inbred lines of one species only. In many cases, the proteins have completely different mutations that result in the same outcome: loss of function. A clear example is found in the tyrosinase protein, in which loss of function results in albinism. Several inbred lines in the mouse have the albino phenotype, related to an amino acid substitution at position 103 (cysteine loss: C103S, predicted deleterious: PROVEAN score = -9.74) [20]. All mouse lines with a mutation in the *Tyr* gene have this variant and are albino. In the rat there are also many inbred strains with an albino phenotype and the *Tyr* gene is also mutated in those. However, none of the rat strains have a C103S variant and indeed all match the rat reference sequence at that position, a cysteine (supplemental figure 1). Instead, the albino rat strains share a different mutation at position 299 (R299H, predicted deleterious: PROVEAN score = -3.21). This variant at position 299 is rat-specific, and was shown to cause albinism in F344 in a previous study [21], and all mouse strains have the same sequence here as the rat reference strain: an arginine (supplemental figure 2).

Position equivalent variants

For the 619 variants that occur at equivalent positions in mouse and rat, we found four different types (Figure 3). All variant possibilities were compared, so for example if a rat variant has 2 mouse variants (in different strains) at the same position, it is included twice. There are a total of 120 variants (group i) that are completely identical in occurrence in mouse and rat, meaning that the gave the same reference amino acid (AA) that is changed to the same alternate residue is some strains (e.g. a V to D in both). One such example is the *Uspl1* gene, which has a shared position between rat and mouse: an A164L substitution but is not predicted to be deleterious in rat or mouse. A little less variant positions,

i.e. 109 (group ii), have the same reference AA but a different alternative strain specific AA (e.g. V to D in mat and V to M in mouse). A practical example is the *Sun2* gene, which has a mutation resulting in a the replacement of an alanine at position 285. In mouse this results in a threonine (MOLF/EiJ; score: -2.16) while in rat a glycine is found instead (ACI/EurMcwi; score: -1.82). The smallest group (group iii) has *convergent variants,* i.e. where the sites have a different reference AA, but the mutation results in the same AA (e.g. D to V in rat and A to V in mouse) such as found for the *Mylk* gene where in mouse there is a V at A substitution at position 465 (MOLF/EiJ,; score: 1.62) and the rat equivalent of this position (455) shows a T to A change (SBH/Yg; score: 1.13). Finally, the largest group (group iv) of 334 variants shows no relation between mouse and rat for position equivalent variants.

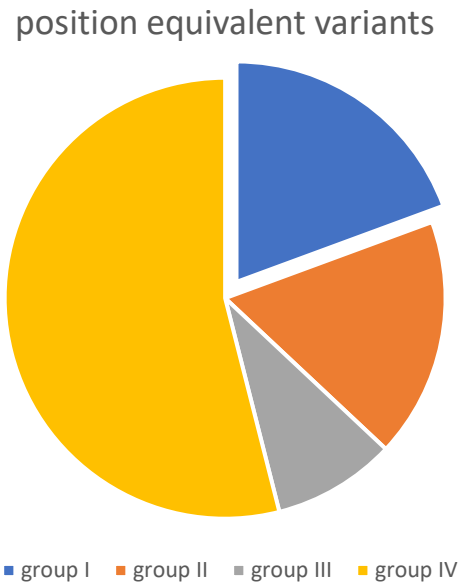

Figure 3. Proportion of the different groups of the 619 shared variant positions. Group I, the set of identical inbred line changes in both mouse and rat, is highlighted. Groups, I (identical variants in rat/mouse), II (same position but different substitution in rat and mouse), III (different reference AA at the position, but changed to identical alternative) and IV (mouse and rat reference as well as variants are different for the position)

This total of 619 shared positions corresponds with 464 distinct protein sequences in each species. A gene set enrichment analysis for pathways and gene ontology (GO) using Metascape terms shows an over-representation of primarily immune process related functions (Figure 4).

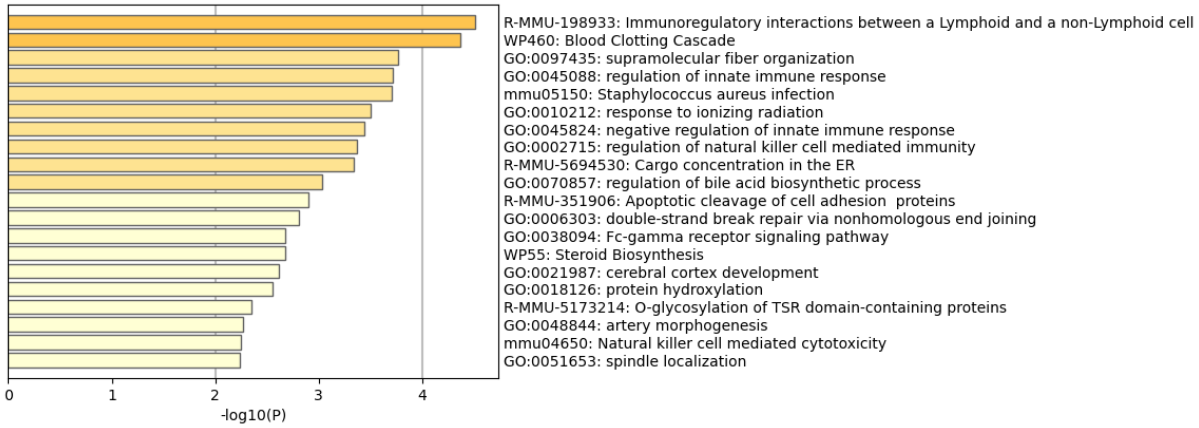

Figure 4 gene set enrichment analysis for overrepresented pathways and functions for the 464 genes containing position conserved variants in mouse and rat inbred strains. The top 20 pathways and functions are enriched for immune related processes, such as signaling and also blood clotting and to a lesser extent DNA damage/repair. Gene set enrichment was performed using metascape and the annotation of the mouse genes of the ortholog pairs.

These functions may be related to the environmental pressures present in these rats and mice. Laboratory animals are known to have a somewhat altered immune function compared to their wild counterparts, mainly due to lower exposure to pathogens [22, 23] when kept in a protected environment such as an specific pathogen free animal house. This may result in a convergent evolution of genes in immune processes over many generations.

In addition, especially the variants in group i are potentially very useful: these positions (both the reference and the alternative AA) have been conserved through evolution or occurred independently twice. If these variants have a PROVEAN score less than -2.5 and are predicted to have a negative effect on protein function. This -2.5 is the cut-off point below which a variant is considered to be deleterious for protein function, if accepting the published 80% balanced accuracy (same true positive and true negative rate) as reported in the PROVEAN manual [24]. It is possible that other AA substitutions have a more severe, intolerable, effect.

Conclusion

We have performed an in-depth comparison of our previously published dataset of protein coding variants in mice (Mousepost) and rats (Ratpost). We show that while the set of transcripts that is affected by a variant in both species shows a very large overlap, also in part that due the fact the

majority of transcripts will have one or more variants if a sufficient number of strains are sequenced, this is not the case for individual variants, which show only minimal overlap. Overall the ”Ratpost” and Mousepost databases, as well as the mouse and rat model organisms are mainly complementary where protein coding variants are concerned. This large repository of natural variations may serve as a useful tool to select a specific inbred strain and species for a (human) disease model.

Materials & methods

Coding sequence variants. We obtained coding sequence variation data from the ”Mousepost” and ”Ratpost” databases. These data were derived from the mouse genomes project for mouse inbred lines and from the rat genome database for rat inbred strains. The collection of complete protein sequences from the construction of ”Ratpost” and ”Mousepost” databases was used in this analysis [15].

Orthologous genes and transcripts. Information concerning orthologous mouse-rat relationships was downloaded from the Ensembl Biomart webtool [25]. The data filtering settings were specified as protein coding, orthologous mouse genes only with the ‘mouse homology type’ attribute added. The resulting datafile was filtered to obtain a set of one-to-one relationships.

Sequence alignment and position conversion. We made use of the python scripting language (v3.8) for all steps of the analysis, global pairwise sequence alignments were performed using the Biopython [26] align module using the Blosum62 substitution matrix and gap opening penalty of 10 and extension penalty of 0.5. Alignments results were kept in memory and processed into a lookup table of mouse to rat (and rat to mouse) position matches. The ”Mousepost” and ”Ratpost” mysql databases were queried for all variants in the transcript under the from “reference_AA position alternative_AA”. Variant position ware compared based on the position lookup table and exact AA at the position.

Gene set enrichment. Enrichment analysis on the set of genes from the 619 shared positions was performed using Metascape [27].

References

1. *EUR-Lex - 52020DC0016 - EN - EUR-Lex.*
2. Philipeaux, J., *Note sur l'exstirpation des capsules surrenales chez les rats albinos (Mus rattus).* Compt. Rend. Hebd. Seances Acad. Sci, 1856. **43**: p. 904-906.
3. Modlinska, K. and W. Pisula, *The Norway rat, from an obnoxious pest to a laboratory pet.* eLife, 2020. **9**.
4. Morse, H., *The laboratory mouse-a historical perspective.* The mouse in biomedical research, 1981: p. 1-16.
5. Bergman, I., et al., *Comparison of in vitro antibody-targeted cytotoxicity using mouse, rat and human effectors.* Cancer immunology, immunotherapy : CII, 2000. **49**(4-5): p. 259-66.
6. Ellenbroek, B. and J. Youn, *Rodent models in neuroscience research: is it a rat race?* Disease models & mechanisms, 2016. **9**(10): p. 1079-1087.
7. Mouse Genome Sequencing, C., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
8. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.
9. Keane, T.M., et al., *Mouse genomic variation and its effect on phenotypes and gene regulation.* Nature, 2011. **477**(7364): p. 289-94.
10. Sharp, P.E. and J.S. Villano, *The laboratory rat*. 2nd ed. The laboratory animal pocket reference series. 2013, Boca Raton, FL: CRC Press. xxi, 377 p.
11. Atanur, S.S., et al., *Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat.* Cell, 2013. **154**(3): p. 691-703.
12. Hermsen, R., et al., *Genomic landscape of rat strain and substrain variation.* BMC genomics, 2015. **16**: p. 357.
13. Smith, J.R., et al., *The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform.* Nucleic Acids Research, 2020. **48**(D1): p. D731-D742.
14. Timmermans, S., M. Van Montagu, and C. Libert, *Complete overview of protein-inactivating sequence variations in 36 sequenced mouse inbred strains.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(34): p. 9158-9163.
15. Timmermans, S. and C. Libert, *Ratpost: a searchable database of protein-inactivating sequence variations in 40 sequenced rat-inbred strains.* Mammalian genome : official journal of the International Mammalian Genome Society, 2021. **32**(1): p. 1-11.
16. Beck, J.A., et al., *Genealogies of mouse inbred strains.* Nature Genetics, 2000. **24**(1): p. 23-+.
17. Ibarra-Soria, X., et al., *Variation in olfactory neuron repertoires is genetically controlled and environmentally modulated.* Elife, 2017. **6**.
18. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic Acids Research, 2018. **46**(D1): p. D1062-D1067.
19. Poltorak, A., et al., *Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene.* Science (New York, N.Y.), 1998. **282**(5396): p. 2085-8.
20. Yokoyama, T., et al., *Conserved Cysteine to Serine Mutation in Tyrosinase Is Responsible for the Classical Albino Mutation in Laboratory Mice.* Nucleic Acids Research, 1990. **18**(24): p. 7293-7298.
21. Lu, B.S., et al., *Generation of rat mutants using a coat color-tagged Sleeping Beauty transposon system.* Mammalian Genome, 2007. **18**(5): p. 338-346.
22. Viney, M., L. Lazarou, and S. Abolins, *The laboratory mouse and wild immunology.* Parasite immunology, 2015. **37**(5): p. 267-73.
23. Yeung, F., et al., *Altered Immunity of Laboratory Mice in the Natural Environment Is Associated with Fungal Colonization.* Cell Host & Microbe, 2020. **27**(5): p. 809-+.

24. Choi, Y., et al., *Predicting the functional effect of amino acid substitutions and indels.* PloS one, 2012. **7**(10): p. e46688.
25. Yates, A.D., et al., *Ensembl 2020.* Nucleic acids research, 2020. **48**(D1): p. D682-D688.
26. Cock, P.J.A., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.
27. Zhou, Y.Y., et al., *Metascape provides a biologist-oriented resource for the analysis of systems-level datasets.* Nature Communications, 2019. **10**.