Abschlussbericht zu den in den wettbewerblichen Verfahren
der Leibniz-Gemeinschaft geförderten Vorhaben

Titel des Vorhabens: Identifying genomic loci underlying mammalian phenotypic variability using *Forward Genomics* with Semantic Phenotypes

Projektnummer/Aktenzeichen: SAW-2016-SGN-2

Executive Summary

Identifying the genomic loci underlying mammalian phenotypic variability is a big challenge in bioscience with strong implications for biomedicine. Analyses of model organisms allow analysing the effect of protein coding and non-coding genomic loci on phenotypic traits. However, this insight does not necessarily allow conclusions about genomic changes that are actually associated to phenotype evolution. Consequently, the concept of *Forward Genomics* was developed to explore genome-phenotype associations in an evolutionary context. This approach searches for common convergent genomic changes (e.g., gene loss) for convergent mammalian phenotypes. Given that this approach was at the proof-of-principle-stage by the time of application (2015/16), the project team performed the first large scale screen for genomic loci underlying phenotype evolution using *Forward Genomics*.

One challenge was establishing a matrix reporting on traits that were associated with convergent phenotype evolution. This resulted in a novel data source specifically designed for the use in comparative genomics, i.e., MaTrics (**Ma**mmalian **Tr**aits for Comparative Genom**ics**). It is hosted by Morph·D·Base and uses the matrix tool that was improved as one outcome of the project. Recording phenotypes revealed substantial knowledge gaps and this motivated launching research, e.g., on comparative morphology of tails, teeth, gall bladder or the vomeronasal system (VNS). The application of *Forward Genomics* to phenotype data allowed the discovery of numerous associations between protein coding genes and phenotypic traits associated with adaptations to aquatic environments, diet, VNS loss, tooth enamel or testicular descent. Furthermore, we developed tools for comparative genomics and generated large comparative genomic resources (e.g., mammalian genome alignments). Integrative research between different project modules focussed on mammalian nutrition. For instance, we analysed diet composition of extant and extinct species and were able to demonstrate that convergent gene losses can be closely associated to the level of fat or toxin consumption. Finally, the project team successfully established methods (ATAC-seq) and mouse models (PHGR1 knock-out) to experimentally validate the function of genomic loci that have been identified by *Forward Genomics* and project associated genome wide association studies. The expertise developed during the course of this study led to establishing national and international collaborations and follow-up research activities covering aspects such as mammalian nutrition, olfaction, and VNS pathology. The project was performed as a collaboration between 8 German institutions involving 26 scientists (13 women, 13 men). The project results are presented in 18 peer-reviewed publications, one dataset (MaTrics; accessible at https://www.morphdbase.de/?MaTrics-Mx288-v1) and three master theses. In addition, scientific results were presented at conferences and workshops (21 talks, 5 posters). Project participants organized the 93rd annual meeting of German Society for Mammalian Biology (2019, Dresden) and the project outcome was presented in key note talks and oral presentations.

Overall, the project was the first large scale application of *Forward Genomics* and could substantially contribute to identifying genomic loci underlying mammalian phenotype variability. This novel scientific approach will in short term have the potential to unravel more genome-phenotype associations (e.g., associated to physiological traits) with implications on evolutionary biology and biomedicine contribution. We argue that *Forward Genomics* is applicable to research also in non-mammalian taxa such as fishes, birds and many invertebrates.

1. Realisation of goals and milestones

The overall aim of the project was to explore genomic loci underlying phenotype variability in mammals. This was done using *Forward Genomics*, a novel concept that was at the proof-of-principle-stage by the time of application (2015/16). This methodology compares genome alignments with trait matrices in order to detect associations between convergent genomic changes (e.g., gene loss) and convergent phenotypes. In contrast to investigating model organisms, this allows identifying candidate genomic loci that played a key role in phenotype evolution and, therefore, may become good candidates for biomedical research. The first large scale application of *Forward Genomics* required reaching seven milestones (modules M1 - M7). Being aware that phenotype information was not publically available in a format required for *Forward Genomics* (e.g., encoded as discrete categories), activities in modules M1 and M2 successfully provided trait information and contributed to closing knowledge gaps on mammalian phenotypes. This information was made publically available in Morph·D·Base (MDB) thanks to implementing novel features into the matrix tool of this data repository (module M3). Activities in module M4 applied *Forward Genomics* and identified candidate genomic loci that explain phenotype variability, e.g., associated to nutrition. Integrative research was performed in module M5 to explore the evolutionary significance of gene losses identified in module M4. For instance, we analysed how the consumption of single diet components (fat, toxins) is associated to gene losses during phylogenesis. Finally, project teams associated to modules M6 and M7 successfully established methods (ATAC-seq) for validating non-exonic candidate loci as well as a mouse model to explore the function of a protein coding gene (PHGR1). This gene was identified being preferentially lost in carnivores (module M4) and was associated to diverticular disease (module M7).
It became clear in the early project stages that the functional interpretation of loci detected by *Forward Genomics* (module M4) is challenging. Many characters turned out to be associated with genes where no mouse model was available (e.g., provided by mouse model banks) or that non-exonic loci are the best candidates to explain skeletal phenotype variation (e.g., regulatory elements). This prompted scientists associated to modules M6 and M7 to adjust their scientific agendas. This included for both modules the employment of scientists (PostDoc) instead of PhD students (as originally planned). This was not least necessary to handle the complexity of the task, i.e., establishing ATAC-seq to explore the function of non-exonic loci and to generate a knock-out mouse model (PHGR1). While refocussing on method development/model establishment, experimental validation of candidate genomic loci is now implemented in follow-up projects. The approval of the project in November 2015 was associated with a severe cut in the financial budget. This required reducing personnel and adjusting aims and objectives particularly in modules M3 (Documentation) and M5 (Evolution and Phylogenetics). Regarding module M3, only one PostDoc position was staffed and this scientist focussed on implementing novel features into the matrix tool of MDB. Consequently, issues related to *Semantic Phenotypes* were cancelled from the project agenda. Nevertheless, selected conceptual aspects related to *Semantic Phenotypes* were addressed by one of the Principle Investigators (PI) (Lars Vogt). Regarding module M5, a scientific trainee was part time employed instead of a PhD student. Consequently, the scientific agenda of module M5 was substantially re-designed and the workload was reduced to a level basically manageable by the PI (Heiko Stuckas) alone. The first new focus was on experimentally re-analysing candidate genomic loci in selected mammalian species for which no high-quality genomic data were available at the time of the respective study. The second new focus was on leading integrative research projects focussing on the evolutionary significance of gene losses in the light of gene function and phenotype.

2. Activities and problems

Research activities associated to modules M1 and M2 were dedicated to phenotype recording and research on phenotype variability. It was based in institutions hosting libraries, collections and laboratories equipped for phenotype analyses (e.g., microscopy, photo-documentation; Senckenberg Gesellschaft für Naturforschung (SGN), Leibniz-Institute for Zoo- and Wildlife Research (IZW), Museum für Naturkunde (MfN)). Both modules provided the required phenotype information (quality and number) and, therefore, guaranteed performing genomic analyses (*Forward Genomics*). The implementation (programming) of novel features into MDB was performed within the framework of module M3 at Zoologisches Forschungsmuseum Alexander König (ZFMK) and University Bonn. These institutions developed and host MDB and hence provide the required infrastructure to perform this work. These improvements were necessary for publication of phenotype information and their use in *Forward Genomics*. Scientists associated to module M4 identified genomic loci underlying phenotype variability (*Forward Genomics*) by analysing genome alignments and trait matrices (e.g., resulting from modules M1-3). This work was performed at Max Planck Institute (MPI) for Cell Biology and Genetics (MPI-CBG) and MPI for the Physics of Complex Systems (MPI-PKS) and used the computational facilities of these institutions. It was

necessary to finally address the overall question of this project. Evolutionary analyses in module M5 required using laboratory facilities of SGN to perform target gene analyses on museum samples (first new focus). Furthermore, computational facilities at SGN allowed performing computations associated to evolutionary analyses (second new focus). This work contributed to identifying candidate genomic loci (together with module M4) and to explore the evolutionary significance of gene losses, e.g., in the context of nutrition. Modules M6 and M7 layed the basis for experimental validation of non-exonic candidate loci (ATAC-seq; Center for Regenerative Therapies Dresden (CRTD), TU Dresden (Germany) and Institute of Molecular Pathology (IMP) Vienna (Austria)) and protein coding candidate genes (knock-out (KO) mouse; CRTD and University Hospital at TU Dresden (UKD)). Experimental validation of candidate genomic loci is necessary in cases were functional knowledge is missing. Experimental approaches (ATAC-seq) and target genes (KO mouse) were selected in close collaboration (basically team members associated to modules M4, M6, M7) and with view on the scientific question.
None of the results obtained by project partners in modules M1-M7 were obtained by other scientists worldwide.
None of the strategies had to be changed due to novel developments in any field of science.

3. Results and achievements

Results of the project demonstrate that *Forward Genomics* is a scientific concept that is best suited to discover genomic causes (e.g., gene loss) underlying phenotypic variability in mammals. This project has shown that gene loss (reduction of the genome's functional repertoire) is an evolutionary event that contributes to phenotypic variability among mammals. We demonstrated that gene loss is associated with adaptations to an aquatic environment, dietary adaptations, the loss of the vomeronasal system (VNS), tooth enamel or testicular descent in mammals. The application of *Forward Genomics* depends on high quality phenotype data. Therefore, the project team established MaTrics (**Ma**mmalian **Tr**aits for comparative Genom**ics**). This is a new data source made publically available as MaTrics version 1.0 (released in January 2021; https://www.morphdbase.de/?MaTrics-Mx288-v1) and is implemented in the data repository Morph·D·Base (MDB, www.morphdbase.de). It has unique key features such as phenotype description by discrete categories that are linked to at least one supporting reference (e.g., literature, collection-ID) to make the data entries revisable. By establishing MaTrics, we provide a concept for a data source specifically tailored for applications in comparative genomics. This is for instance because phenotype documentations can be automatically obtained in machine actionable formats (e.g., trait matrix in NEXUS or CSV file format). This is one out of many new features implemented into the MDB matrix module during the course of the project. Providing phenotype information specifically tailored for use in comparative genomics fills a research gap and brings scientific institutions hosting collections (e.g., SGN, MfN, ZFMK, IZW) in a new strategic position. This prediction was already made in our application. The strategic value of this data source is for instance illustrated by the fact that MaTrics opened the door for collaborative links to the Zoonomia Consortium (Broad Institute of MIT and Harvard, Cambridge, MA USA; https://zoonomiaproject.org/meet-the-team/) (see below for details). We further unravelled substantial knowledge gaps on phenotypic traits for many mammalian species and lineages. Closing these gaps will be essential not only to solve questions related to genome-phenotype associations using *Forward Genomics*. For instance, comparative studies on rodent tails provided new insight into organismic evolution. Similarly, the project team was motivated to launch research on dentition/teeth, skull, extrahepatic biliary tract pathologies, gall bladder or the VNS. By establishing methods (ATAC-seq) and a knock-out mouse model (PHGR1 KO model; see below for details), the project team lays the basis for an experimental validation of candidate genomic loci. For instance, ATAC-seq can contribute to the experimental validation of non-exonic candidate genomic loci (e.g., regulatory elements). Thus, this method has a great potential for exploring the genomic causes of variability seen in mammalian dentition (126 dentition traits are already recorded in MaTrics). Finally, we would like to highlight that integrative research activities brought together the expertise of different scientific disciplines (modules M1-M7). This resulted in a research focus on mammalian nutrition**.** We provide a list of candidate genes that are preferentially lost in different mammalian lineages depending on their nutrition strategy, i.e., herbivory and carnivory. Subsequently, the project team demonstrated that some of these gene losses are associated with the overall content of specific diet components instead of a strictly carnivore or herbivore nutrition strategy. This refers to components such as fat (gene encoding the lipase inhibitor PNLIPRP1) or toxins (gene encoding the xenobiotic receptor NR1I3). This study is based on comprehensive research on diet composition in different mammalian species documented and made accessible in MaTrics. Furthermore, this focus on mammalian nutrition revealed that the protein coding gene PHGR1 is not only preferentially lost in carnivorous mammals but also associated with the diverticular disease in humans. This is why the PHGR1 KO mouse model (see above) is considered having a great scientific value. In fact, it bears the potential to understand PHGR1 gene function from different perspectives: the wildtype gene function in mammals and the function as candidate disease gene in humans.

The scientific output of the project is first of all documented by 18 peer-reviewed publications in international journals. This is complemented by conference- and workshop contributions (21 oral presentations, 5 posters). Furthermore, project participants were among the organizers of the 93rd Annual Meeting of Deutsche Gesellschaft für Säugetierkunde, e. V. (DGS; German Society for Mammalian Biology) in September 2019. The session on "Phenotypes and morphotypes: assessment, function, development, and evolution" gave the opportunity to report on advances made during the course of the project (two key note lectures, three additional oral presentations). Finally, the project gave three master students the chance to write their theses.
One precondition for routine use of *Forward Genomics* is the availability of phenotype information, e.g., in a format implemented in MaTrics as described above. This is planned to be done by the MaTrics-Consortium (see also https://www.senckenberg.de/de/institute/senckenberg-naturhistorische-sammlungen-dresden/museum-fuer-tierkunde/dd-sekt-mammalogie/matrics-consortium/). This is a group of scientist from SGN, MfN, and IZW (organismic expertise) as well as ZFMK and TIB Leibniz Information Centre for Science and Technology (implementation of Data in MorpDBase). Results highlighted above are also the basis for ongoing research agendas and grant applications (e.g., DFG). One focus is on the continuation of the mammalian nutrition focus. Furthermore, team members (SGN) established a collaboration with Zoonomia Consortium (see above) to perform a pioneering investigation on the association between the number of olfactory receptors (OR) in the genomes of more than 250 placental mammals (representing all orders) and the olfactory turbinal count, the surface area of olfactory turbinals and epithelium, and olfactory bulb size. Similarly, national and international collaborations aim at identifying genomic signals and embryological patterns that might explain the differences observed in short-tailed mammals as well as the mechanics that lead to these structures. Finally, follow-up research on VNS morphology is initiated based on a successful grant application (MfN).

4. Equal opportunities and Internationalization

The guarantee of equal gender opportunities was an agreement by the time of application and is reflected by the choice of the leading scientist team (3 women, 4 men). We were successful in forming a final team of scientist (PI and hired positions: Phd, PostDoc, students/trainees) consisting of 13 women and 13 men. We would like to emphasize that one project partner (IZW) employed a deaf person as PhD student. We acknowledge that IZW provided additional money to finance interpreters that guaranteed communication, e.g., during daily business but also in workshops and at conferences. This initiative made substantial contributions to phenotype recording and research on gall bladder possible. The project was performed as a collaboration between 8 German institutions involving 26 scientists; most of them hold a German citizenship. One PI (Prof. Elly Tanaka) is US American citizen and moved to IMP Vienna (Austria).

5. Structures and collaborations

The project was performed as collaboration between team members located at 8 partner institutions (SGN, IZW, MPI-CBG, MPI-PKS, MfN, ZFMK, CRTD, UKD, University Bonn). No additional collaborations were established during the course of this project. A kick-off meeting brought mainly PI's of all partner institutions together (May 2016). Afterwards, regular meeting in person or via video calls concentrated on specific topics, e.g., Introduction into MDB (scientist associated to modules M1, M2, M3, M5), phenotyping and application of *Forward Genomics* (scientist associated to modules M1, M2, M4, M5), Evolution (scientist associated to modules M1, M2, M4, M5) and experimental validation (scientist associated to modules M4, M6, M7). These project specific meetings guaranteed the coordination of overlapping activities. The end of the project duration (30.06.2020) fall into the time of the arising Covid-19-pandemia in Germany. This was associated with strong restrictions which made a final meeting impossible.

6. Quality management

Because as all partner institutions are located in Germany, PI's agreed on following the DFG-rules to guarantee good scientific practice. The outcome that defines the scientific quality of the project refers to i) publications, ii) phenotype information in MaTrics, and iii) experimental models. All 18 scientific publications passed the peer-review as well as the journal specificities on good scientific practice and data availability etc. (see list of publications). MaTrics version 1.0 (released in January 2021; https://www.morphdbase.de/?MaTrics-Mx288-v1) is available online including the documentation (Wagner et al., 2021; https://doi.org/10.20363/mdb.ref-5293). Establishing experimental models included experiments on animals. First, establishing ATAC-seq required collecting tissue from animals in this procedure is legally covered by permissions given to CRTD and IMP represented by Prof. Elly Tanaka (lead of module M6). Generating the knock-out mouse is based on the

permission given by Landesdirektion Sachsen (TVA ID: TVV 40/2021) to CRTD, UKD and Prof. Jochen Hampe (lead of module M7).

7. Additional Resources

The total number of 13 PI's were involved based on contracts with their institutions. We estimate that PI's spent on average 20% of their time in project contributions over the period of 36 month (excluding cost neutral extension time). Thus, PI's contributed approx. 94 month (Personenmonate). The approval of the project in November 2015 was associated with a severe cut in the financial budget by approx. 32%. This was compensated by hiring less personnel (see above) and institutions provided consumables, travel expenses, and workshop budgets equivalent to approx. 215 000 Euro.

8. Outlook

The project was the first large scale application of *Forward Genomics*, a concept that was at early stages of its development by the time of application (2015/16). By applying this concept, we identified a number of genomic loci underlying mammalian phenotype variability and showed that gene loss is a factor contributing to mammalian evolution. Consequently, we argue that *Forward Genomics* will have the potential to unravel more genome-phenotype associations (e.g., involving gene losses). This will include research of adaptive traits (examples are provided as project results). In this particular context, we predict the identification of genomic loci that allow an understanding of how species can adapt to globally changing environmental conditions. At short term, the focus will be on identifying protein coding genes associated to physiological traits. The intermediate term perspective will be on identifying also non-exonic loci (e.g., regulatory elements) associated to phenotypes such as skeletal variability (e.g., dentition). We argue that *Forward Genomics* is applicable to research also on non-mammalian taxa such as fishes, birds and many invertebrates.