In silico candidate variant and gene identification using inbred mouse strains

Matthias Munz Equal first author, 1 , **Mohammad Khodaygani** Equal first author, 1 , **Zouhair Aherrahrou** 2 , **Hauke Busch** Corresp., 1 , **Inken Wohlers** Corresp. 1

[1] Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology and Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany
[2] Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany

Corresponding Authors: Hauke Busch, Inken Wohlers
Email address: Hauke.Busch@uni-luebeck.de, Inken.Wohlers@uni-luebeck.de

Mice are the most widely used animal model to study genotype to phenotype relationships. Inbred mice are genetically identical, which eliminates genetic heterogeneity and makes them particularly useful for genetic studies. Many different strains have been bred over decades and a vast amount of phenotypic data has been generated. In addition, lately, also whole genome sequencing-based genome-wide genotype data for many widely used inbred strains has been released. Here, we present an approach for *in silico* fine mapping that uses genotypic data of 37 inbred mouse strains together with phenotypic data provided by the user to propose candidate variants and genes for the phenotype under study. Public genome-wide genotype data covering more than 74 million variant sites is queried efficiently in real-time to provide those variants that are compatible with the observed phenotype differences between strains. Variants can be filtered by molecular consequences and by corresponding molecular impact. Candidate gene lists can be generated from variant lists on the fly. Fine mapping together with annotation or filtering of results is provided in a Bioconductor package called MouseFM. For albinism, MouseFM reports only one variant allele of moderate or high molecular impact that only albino mice share: a missense variant in the *Tyr* gene, reported previously to be causal for this phenotype. Performing *in silico* fine mapping for interfrontal bone formation in mice using four strains with and five strains without interfrontal bone results in 12 genes. Of these, three are related to skull shaping abnormality. Finally performing fine mapping for dystrophic cardiac calcification by comparing 8 strains showing the phenotype with 8 strains lacking it, we identify only one moderate impact variant in the known causal gene *Abcc6*. In summary, this illustrates the benefit of using MouseFM for candidate variant and gene identification.

In silico candidate variant and gene identification using inbred mouse strains

Matthias Munz[1*], Mohammad Khodaygani[1*], Zouhair Aherrahrou[2], Hauke Busch[1#], and Inken Wohlers[1#]

[1]**Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology and Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany**
[2]**Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany**
[*,#]**These authors contributed equally**

Corresponding author:
Hauke Busch and Inken Wohlers

Email address: hauke.busch@uni-luebeck.de and inken.wohlers@uni-luebeck.de

ABSTRACT

Mice are the most widely used animal model to study genotype to phenotype relationships. Inbred mice are genetically identical, which eliminates genetic heterogeneity and makes them particularly useful for genetic studies. Many different strains have been bred over decades and a vast amount of phenotypic data has been generated. In addition, lately, also whole genome sequencing-based genome-wide genotype data for many widely used inbred strains has been released. Here, we present an approach for *in silico* fine mapping that uses genotypic data of 37 inbred mouse strains together with phenotypic data provided by the user to propose candidate variants and genes for the phenotype under study. Public genome-wide genotype data covering more than 74 million variant sites is queried efficiently in real-time to provide those variants that are compatible with the observed phenotype differences between strains. Variants can be filtered by molecular consequences and by corresponding molecular impact. Candidate gene lists can be generated from variant lists on the fly. Fine mapping together with annotation or filtering of results is provided in a Bioconductor package called MouseFM. For albinism, MouseFM reports only one variant allele of moderate or high molecular impact that only albino mice share: a missense variant in the *Tyr* gene, reported previously to be causal for this phenotype. Performing *in silico* fine mapping for interfrontal bone formation in mice using four strains with and five strains without interfrontal bone results in 12 genes. Of these, three are related to skull shaping abnormality. Finally performing fine mapping for dystrophic cardiac calcification by comparing 9 strains showing the phenotype with 8 strains lacking it, we identify only one moderate impact variant in the known causal gene *Abcc6*. In summary, this illustrates the benefit of using MouseFM for candidate variant and gene identification.

INTRODUCTION

Mice are the most widely used animal models in research. Several factors such as small size, low cost of maintain, and fast reproduction as well as sharing disease phenotypes and physiological similarities with human makes them one of the most favourable animal model (Uhl and Warner, 2015). Inbred mouse strains are strains with all mice being genetically identical, i.e. clones, as a result of sibling mating for many generations, which results in eventually identical chromosome copies. When assessing genetic variance between mouse strains, the genome of the most commonly used inbred strain, called black 6 (C57BL/6J) is typically used as reference and variants called with respect to the black 6 mouse genome. For inbred mouse strains, variants are homozygous by design.

Grupe *et al.* in 2001 published impressive results utilizing first genome-wide genetic data for *in silico* fine mapping of complex traits, "reducing the time required for analysis of such [inbred mouse] models from many months down to milliseconds" (Grupe et al., 2001). Darvasi commented on this paper that in his opinion, the benefit of *in silico* fine mapping lies in the analysis of monogenic traits and in informing researchers prior to initiating traditional breeding-based studies. In 2007, with Cervino *et al.*, he suggested to combine *in silico* mapping with expression information for gene prioritization using

20,000 and 240,000 common variants, respectively (Cervino et al., 2007). Although genetic data improved incredibly since then – now all genetic variation between all commonly used inbred strains is known at base pair resolution (Doran et al., 2016) (Keane et al., 2011) – to the best of our knowledge, the idea of *in silico* fine mapping using inbred mouse strains has not been picked up again since then.

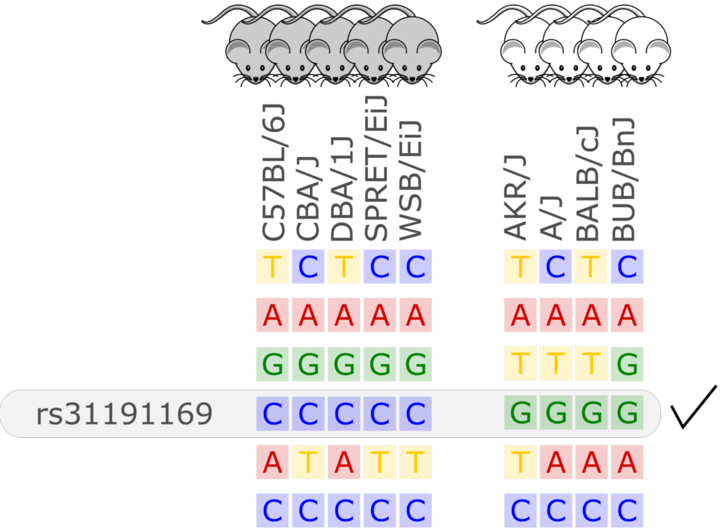
At the same time, in the last years huge amounts of mouse phenotype data were generated, often in collaborative efforts and systematically for many mouse strains. Examples are phenotyping undertaken by the International Mouse Phenotyping Consortium (IMPC) (Dickinson et al., 2016)(Meehan et al., 2017) or lately also the phenotyping of the expanded BXD family of mice (Ashbrook et al., 2019). Data are publicly available in resources such as the mouse phenome database (MPD) (Bogue et al., 2018) (`https://www.mousephenotype.org`) or the IMPC's website (Dickinson et al., 2016) (`https://phenome.jax.org`). Other websites such as Mouse Genome Informatics (MGI) (`http://www.informatics.jax.org`) or GeneNetwork (Mulligan et al., 2017) (`https://www.genenetwork.org`) also house phenotype data together with web browser-based functionality to investigate genotype-phenotype relationships.

Several of the aforementioned resources allow to interactively query genotypes for user-selected inbred mouse strains for input genes or genetic regions. None of them though provides the functionality to extract genome-wide all variants that are different between two user-specified groups of inbred mouse strains. Such information can be used for *in silico* fine mapping and for the identification of candidate genes and variants underlying a phenotypic trait. Further, such a catalog of genetic differences between groups of strains is very useful prior to designing mouse breeding-based experiments e.g. for the identification or fine mapping of quantitative trait loci (QTL).

METHODS

Fine mapping approach

Unlike previous approaches for *in silico* fine mapping, here we are using whole genome sequencing-based variant data and thus information on all single nucleotide variation present between inbred strains. Due to the completeness of this variant data, we do not need to perform any statistical aggregation of variant data over genetic loci, but simply report all variant sites with different alleles between two groups of inbred strains. That is, we report all variant sites with alleles compatible with the observed phenotype difference, see Figure 1 for an illustration.



	C57BL/6J	CBA/J	DBA/1J	SPRET/EiJ	WSB/EiJ	AKR/J	A/J	BALB/cJ	BUB/BnJ	
	T	C	T	C	C	T	C	T	C	
	A	A	A	A	A	A	A	A	A	
	G	G	G	G	G	T	T	T	G	
rs31191169	C	C	C	C	C	G	G	G	G	✓
	A	T	A	T	T	T	A	A	A	
	C	C	C	C	C	C	C	C	C	

Figure 1. Illustration of the *in silico* fine mapping approach. Every row represents a variant site and every column one inbred mouse strain. In this example, the phenotype is albinism and four strains are albinos and 5 are not. Displayed are six variants, but only one variant, rs31191169, has consistently different alleles between the albino and the other mice (G allele is here linked to albinism). With option thr2=1 in the MouseFM package, one discordant strain would be allowed in the second strain group and the variant in the row above rs31191169 would also be returned.

In the case of a binary phenotype caused by a single variant, this causal variant is one of the variants that has a different allele in those strains showing the phenotype compared to those strains lacking the phenotype. This is the case for example for albinism and its underlying causal variant rs31191169, used in Figure 1 for illustration and discussed later in detail.

This *in silico* fine mapping approach can reduce the number of variants to a much smaller set of variants that are compatible with a phenotype. The more inbred strains are phenotyped and used for comparison, the more variants can be discarded because they are not compatible with the observed phenotypic difference.

In the case of a quantitative phenotype, the fine mapping can be performed in two ways. The first option is to obtain genetic differences between strains showing the most extreme phenotypes. The second option is binarization of the phenotype by applying a cutoff. Since in these cases allele differences of variants affecting the trait may not be fully compatible with an artificially binarized phenotype, fine mapping is provided with an option that allows alleles of a certain number of strains to be incompatible with the phenotype, see Figure 1 for an example.

Variant data

The database used by this tool was created based on the genetic variants database of the Mouse Genomes Project (`https://www.sanger.ac.uk/science/data/mouse-genomes-project`) of the Wellcome Sanger Institute. It includes whole genome sequencing-based single nucleotide variants of 36 inbred mouse strains which have been compiled by Keane et al. (2011), see `ftp://ftp-mouse.sanger.ac.uk/REL-1502-BAM/sample_accessions.txt` for the accession code and sources. This well designed set of inbred mouse strains for which genome-wide variant data is available includes classical laboratory strains (C3H/HeJ, CBA/J, A/J, AKR/J, DBA/2J, LP/J, BALB/cJ, NZO/HlLtJ, NOD/ShiLtJ), strains extensively used in knockout experiments (129S5SvEvBrd, 129P2/OlaHsd, 129S1/SvImJ, C57BL/6NJ), strains used commonly for a range of diseases (BUB/BnJ, C57BL/10J, C57BR/cdJ, C58/J, DBA/1J, I/LnJ, KK/HiJ, NZB/B1NJ, NZW/LacJ, RF/J, SEA/GnJ, ST/bJ) as well as wild-derived inbred strains from different mouse taxa (CAST/EiJ, PWK/PhJ, WSB/EiJ, SPRET/EiJ, MOLF/EiJ). Genome sequencing, variant identification an characterization of 17 strains was performed by Keane et al. (2011) and of 13 strains by Doran et al. (2016). We downloaded the single nucleotide polymorphism (SNP) VCF file `ftp://ftp-mouse.sanger.ac.uk/current_snps/mgp.v5.merged.snps_all.dbSNP142.vcf.gz`. Overall, it contains 78,767,736 SNPs, of which 74,873,854 are autosomal. The chromosomal positions map to the mouse reference genome assembly GRCm38 which is based on the C57 black 6 inbred mouse strain and by definition has no variant positions.

Low confidence, heterozygous, missing and multiallelic variants vary by strain, in sum they are typically less than 5% of the autosomal variants (Figure 2, Suppl. Table 1. Exceptions are for example the wild-derived inbred strains, for which variant genotypes excluded from the database reach a maximum of 11.5% for SPRET/EiJ. There are four strains that are markedly genetically different from each other and all remaining strains, these are the wild-derived, inbred strains CAST/EiJ, PWK/PhJ, SPRET/EiJ and MOLF/EiJ, see Figure 2A. These four strains also show the highest number of missing and multiallelic genotypes (Figure 2B and Suppl. Table 1).

Database

We re-annotated the source VCF file with Ensembl Variant Effect Predictor (VEP) v100 (McLaren et al., 2016) using a Docker container image (`https://github.com/matmu/vep`). For real-time retrieval of variants compatible with phenotypes under various filtering criteria, the variant data was loaded into a MySQL database. The database consists of a single table with columns for chromosomal locus, the reference SNP cluster ID (rsID), variant consequences based on a controlled vocabulary from the sequence ontology (Eilbeck et al., 2005), the consequence categorization into variant impacts “HIGH”, “MODERATE”, ‘LOW” or “MODIFIER” according to the Ensembl Variation database (Hunt et al., 2018) (see Suppl. Table 2 for details) and the genotypes (NULL = missing, low confidence, heterozygous or consisting of other alleles than reference or most frequent alternative allele; 0 = homozygous for the reference allele, 1 = homozygous for alternative allele). SNPs with exclusively NULL genotypes were not loaded into the database resulting in 74,480,058 autosomal SNVs that were finally added to our database. These have been annotated with overall 120,927,856 consequences, i.e. on average every variant has two annotated consequences. Figure 3 summarizes these consequence annotations stratified by impact; description of consequences and annotation counts are provided in Suppl. Table 2. Most annotations

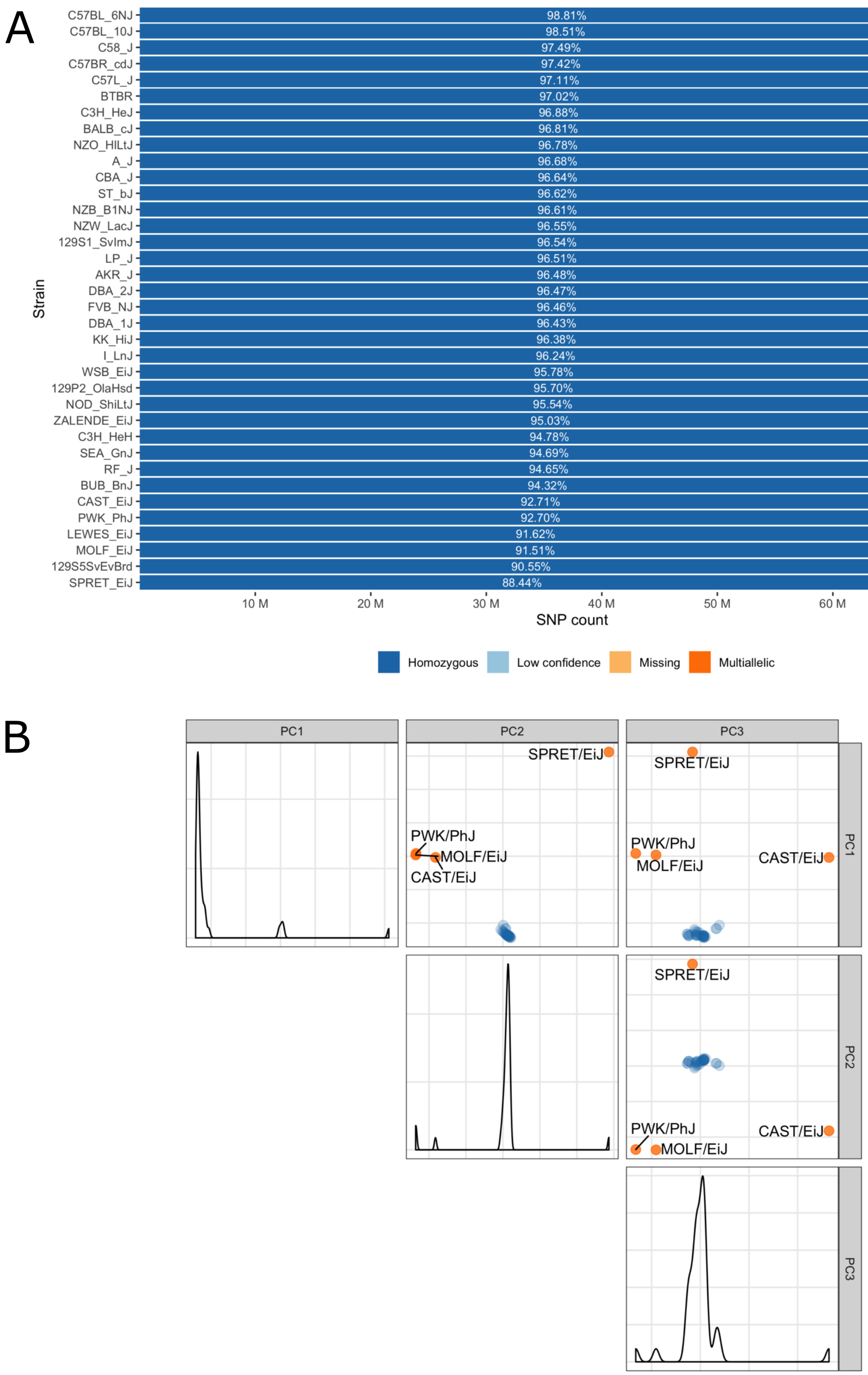


Figure 2. A) Inbred mouse strain autosomal SNP characteristics: The number of homozygous, low confidence, missing and multiallelic genotypes for 36 non-reference strains. For each strain, a SNP was checked for group membership in the order low confidence → missing → multiallelic → homozygous → heterozygous and was assigned to the first matching group. Since no SNP made it to the group with heterozygous genotypes it is not shown in the diagram. B) Principal component analysis shows four outlier inbred strains, CAST/EiJ, PWK/PhJ, SPRET/EiJ and MOLF/EiJ.

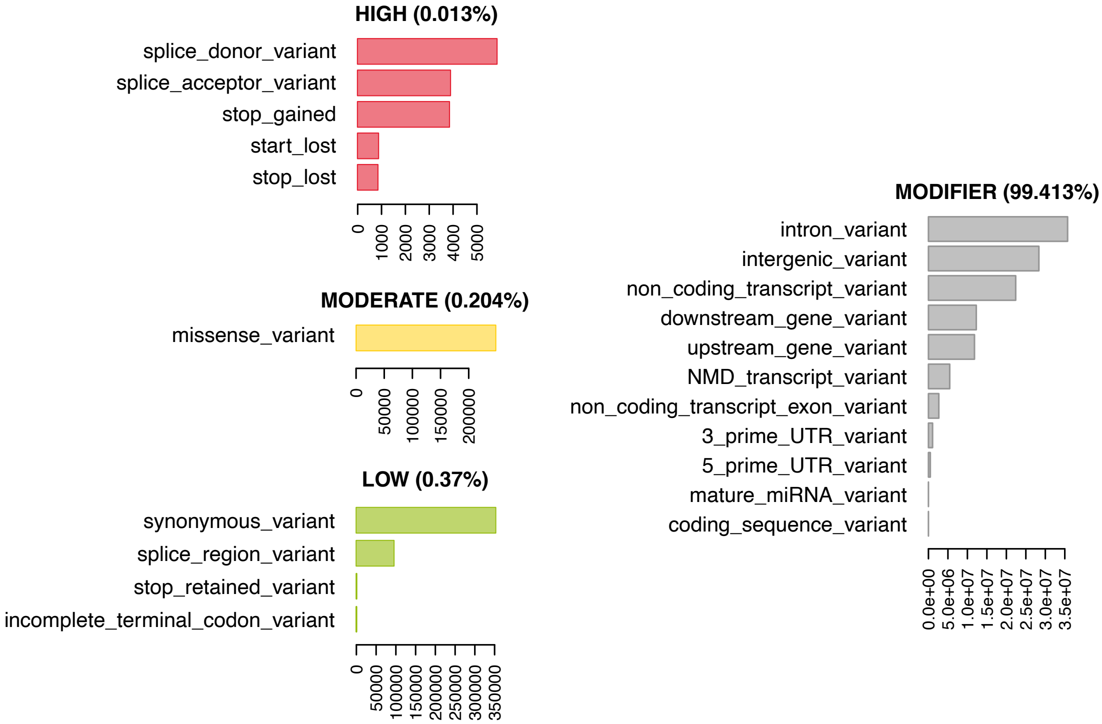


Figure 3. 74,480,058 variants have been annotated with 120,927,856 consequences. Shown here are the number of variants annotated with a given consequence, stratified by consequence impact ("HIGH", "MODERATE",' "LOW", "MODIFIER"). For description of consequence types see Suppl. Table 2. Both impact and consequence can be used for variant prioritization in MouseFM.

belong to impact category "MODIFIER" (99.4%). High impact annotations are rare, because they are typically deleterious (0.013%). Annotation with moderate impact consequences comprise only missense, i.e. protein sequence altering variants contributing 0.204%. Low impact consequences are slightly more often annotated, amounting to 0.37%.

Bioconductor R package MouseFM

Our fine mapping approach was implemented as function `finemap` in the Bioconductor *R* package "MouseFM". Bioconductor is a repository for open software for bioinformatics.

Function `finemap` takes as input two groups of inbred strains and one or more chromosomal regions on the GRCm38 assembly and returns a SNP list for which the homozygous genotypes are discordant between the two groups. Optionally, filters for variant consequence and impacts as well as a threshold for each group to allow for intra-group discordances can be passed. With function `annotate_mouse_genes` the SNP list can further be annotated with overlapping genes. Optionally, flanking regions can be passed.

The `finemap` function queries the genotype data from our backend server while function `annotate_mouse_gene` queries the Ensembl Rest Service (Yates et al., 2014). The repository containing the backend of the MouseFM tool, including the scripts of the ETL (Extract, transform, load) process and the webserver, is available at `https://github.com/matmu/MouseFM-Backend`. Following the repositories' instructions, users may also install the data base and server application on a local server.

RESULTS

As a proof of concept, we applied our *in silico* fine mapping approach on three phenotypes: albinisim, interfrontal bone formation and dystrophic cardiac calcification. Phenotypic data is illustrated in Figure 4.

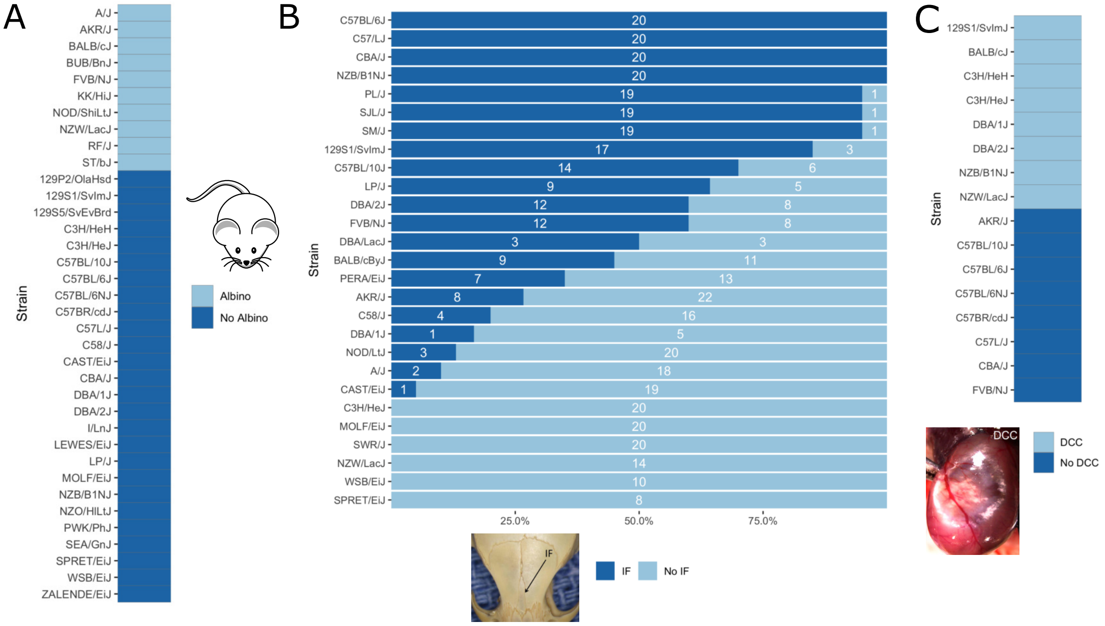


Figure 4. Visualization of mouse phenotypic data for which fine mapping is performed. A) Binary inbred mouse strain phenotype albinism. All or no mice of a strain are albinos; shown here is which strain belongs to which group. B) Quantitative inbred mouse strain phenotype interfrontal bone (IF). Shown is the number of mice of the respective strain having an interfrontal bone (dark blue, IF) and not having an interfrontal bone (light blue, No IF). The interfrontal bone (IF) image is taken from (Zimmerman et al., 2019).

Albinism

Albinism is the absence of pigmentation resulting from a lack of melanin and is well-studied in mice (Beermann et al., 2004). It is a monogenic trait caused by a mutation in the *Tyr* gene (Beermann et al., 2004), which encodes for tyrosinase, an enzyme involved in melanin synthesis. The *Tyr* locus has been used before for the validation of *in silico* fine mapping approaches (Cervino et al., 2007). According to the Jackson Laboratory website (`https://www.jax.org`), 10 of the 37 inbred mouse strains are albinos with a *Tyr*^{c} genotype (`http://www.informatics.jax.org/allele/MGI:1855976`), see Figure 4A.

Our algorithm resulted in only one genetic locus, which includes the *Tyr* gene; only 245 SNPs have different alleles between the albino and non-albino inbred mouse strains, all located from 7:83,244,464 to 7:95,801,713 (GRCm38). When removing SNPs except those of moderate or high impact, only one variant remains. This variant rs31191169 at position 7:87,493,043, with reference allele C and with alternative allele G in the albino strains is the previously described causal missense SNP in the *Tyr* gene, which results in a cysteine to serine amino acid change at position 103 of the tyrosine protein.

Interfrontal bone

Further, we applied our algorithm on the phenotype of interfrontal bone formation, a complex skeletal trait residing between the frontal bones in inbred mice (Figure 4B). In some inbred mouse strains, the interfrontal bone is present or absent in all mice, whereas other strains are polymorphic for this phenotype suggesting that phenotypic plasticity is involved. Phenotypic data related to interfrontal bone has recently been generated by Zimmerman *et al.* (Zimmerman et al., 2019) for 27 inbred mouse strains (Figure 4B). They performed QTL mapping and identified four significant loci on chromosomes 4,7,11 and 14, the same loci for interfrontal bone length and interfrontal bone width. For the genotyping, the authors use the mapping and developmental analysis panel (MMDAP; Partners HealthCare Center for Personalized Genetic Medicine, Cambridge, MA, United States), which contains 748 SNPs.

Of the available interfrontal bone data, we only used inbred strains for which all mice show the same phenotype. This corresponds to four strains with interfrontal bone (C57BL/6J, C57L/J, CBA/J, NZB/B1NJ) and five strains without interfrontal bone (C3H/HEJ, MOLF/EiJ, NZW/LacJ, WSB/EiJ, SPRET/EiJ).

In silico fine mapping resulted in 8,608 SNPs compatible with the observed interfrontal bone phenotype. Of these, 15 showed moderate or high impact on 12 candidate genes, see Table 1. None of the loci identified by us overlaps with the fine mapping results reported by Zimmerman *et al.* Variant rs29393437 is located in the less well described isoform ENSMUST00000131519.1 of *Stac2*, one of two isoforms of this gene. It is is a missense variant, changing arginine (R) to histidine (H) which is at low confidence predicted to be deleterious by SIFT. *Stac2* has been shown to negatively regulate formation of osteoclasts, cells that dissect bone tissue (Jeong et al., 2018). *Phf21* is expressed during ossification of cranial bones in mouse early embryonic stages and has been linked to craniofacial development (Kim et al., 2012). Gene *Abcc6* is linked to abnormal snout skin morphology in mouse and abnormality of the mouth, high palate in human according to MGI.

RSID	Position	Gene
rs32785405	1:36311963	*Arid5a*
rs27384937	2:92330761	*Phf21a*
rs32757904	7:45996764	*Abcc6*
rs32761224	7:46068710	*Nomo1*
rs32763636	7:46081416	*Nomo1*
rs13472312	7:46376829	*Myod1*
rs31674298	7:46443316	*Sergef*
rs31226051	7:49464827	*Nav2*
rs248206089	7:49547983	*Nav2*
rs45995457	9:86586988	*Me1*
rs29393437	11:98040971	*Stac2*
rs29414131	11:98042573	*Stac2*
rs251305478	11:98155926	*Med1*
rs27086373	11:98204403	*Cdk12*
rs27026064	11:98918145	*Cdc6*

Table 1. Moderate and high impact candidate variants and genes for interfrontal bone formation.

Dystrophic cardiac calcification

Physiological calcification takes place in bones, however pathologically calcification may affect the cardiovascular system including vessels and the cardiac tissue. Dystrophic cardiac calcification (DCC) is known as calcium phosphate deposits in necrotic myocardiac tissue independently from plasma calcium and phosphate imbalances. We previously reported the identification of four DCC loci Dyscal1, Dyscalc2, Dyscalc3, and Dyscalc4 on chromosomes 7, 4, 12 and 14, respectively using QTL analysis and composite interval mapping (Ivandic et al., 1996, 2001). The Dyscalc1 was confirmed as major genetic determinant contributing significantly to DCC (Aherrahrou et al., 2004). It spans a 15.2 Mb region on proximal chromosome 7. Finally, chromosome 7 was further refined to a 80 kb region and *Abcc6* was identified as causal gene (Meng et al., 2007; Aherrahrou et al., 2007). In this study we applied our algorithm to previously reported data on 16 mouse inbred strains which were well-characterized for DCC (Aherrahrou et al., 2007). Eight inbred mouse strains were found to be susceptible to DCC (C3H/HeJ, NZW/LacJ, 129S1/SvImJ, C3H/HeH, DBA/1J, DBA/2J, BALB/cJ, NZB/B1NJ) and eight strains were resistant to DCC (CBA/J, FVB/NJ, AKR/J, C57BL/10J, C57BL/6J, C57BL/6NJ, C57BR/cdJ, C57L/J). 2,003 SNPs in 13 genetic loci were fine mapped and found to match the observed DCC phenotype in the tested 16 DCC strains. Of these, 19 SNPs are moderate or high impact variants affecting protein amino acid sequences of 13 genes localized in two chromosomal regions mainly on chromosome 7 (45.6-46.3 Mb) and 11 (102.4-102.6 Mb), see Table 2. The SNP rs32753988 is compatible with the observed phenotype manifestations and affects the previously identified causal gene *Abcc6*. This SNP has a SIFT score of 0.22, the lowest score after two SNPs in gene *Sec1* and one variant in gene *Mamstr*, although SIFT predicts all amino acid changes to be tolerated.

RSID	Position	Gene
rs46174746	7:45538428	*Plekha4*
rs49200743	7:45634990	*Rasip1*
rs32122777	7:45642384	*Mamstr*
rs215144870	7:45679109	*Sec1*
rs45768641	7:45679410	*Sec1*
rs51645617	7:45679423	*Sec1*
rs31997402	7:45725284	*Spaca4*
rs50753342	7:45794044	*Lmtk3*
rs50693551	7:45794821	*Lmtk3*
rs52312062	7:45798406	*Lmtk3*
rs49106901	7:45798469	*Emp3*
rs47934871	7:45918097	*Emp3*
rs32444059	7:45942897	*Ccdc114*
rs32753988	7:45998774	*Abcc6*
rs32778283	7:46219386	*Ush1c*
rs31889971	7:46288929	*Otog*
rs50613184	11:102456258	*Itga2b*
rs27040377	11:102457490	*Itga2b*
rs29383996	11:102605308	*Fzd2*

Table 2. Moderate and high impact candidate variants and genes for dystrophic cardiac calcification.

DISCUSSION & CONCLUSIONS

With MouseFM, we developed a novel tool for *in silico*-based genetic fine mapping exploiting the extremely high homozygosity rate of inbred mouse strains for identifying new candidate SNPs and genes. By including genotype data for 37 inbred mouse strains at a genome-wide scale derived from Next Generation Sequencing, MouseFM clearly outperforms earlier approaches.

By re-analyzing previously published fine mapping studies for albinism and dystrophic cardiac calcificaton, we could show that MouseFM is capable of re-identifying causal SNPs and genes. Re-analyzing a study on interfrontal bone formation (IF), however, did not show any overlap with the regions suggested in the original publication. Reasons might be complex nature of this phenotype and that the causal genetic factors are still largely unknown. With gene *Stac2* we suggest a new candidate gene possibly affecting interfrontal bone formation.

We observe that frequently genetic loci identified by MouseFM fine mapping consist of few or often only a single variant compatible with the phenotype. For example, five of 13 fine mapped DCC loci comprise a single phenotype-pattern compatible variant and 3 loci comprise less than 10 variants. This contradicts the expectation that commonly used mice strains differ by chromosomal segments comprising several or many consecutive variants. Commenly used inbred strains display mosaic genomes with sequences from different subspecific origins Wade et al. (2002) and thus one may expect genomic regions with high SNP rate. Fine mapped loci comprising more phenotype-compatible variants are thus likely more informative for downstream experiments. When allowing no phenotype outlier strain (i.e. thr1=0 and thr2=0), in the case of DCC we identify only six such genetic loci that lend themselves for further experimental fine mapping (chr7:45,327,763-46,308,368 (811 compatible SNVs); chr7:54,894,131-54,974,260 (32 compatible SNVs); chr9:106,456,180-106,576,076 (170 SNVs); chr11:24,453,006-24,568,761 (40 compatible SNVs); chr11:102,320,611-102,607,848 (46 compatible SNVs); chr16:65,577,755-66,821,071 (890 compatible SNVs)).

We show here that *in silico* fine mapping can effectively identify genetic loci compatible with the observed phenotypic differences and prioritize genetic variants and genes for further consideration. This allows for subsequent more targeted approaches towards identification of causal variants and genes using literature, data integration, and lab and animal experiments. MouseFM *in silico* fine mapping provides phenotype-compatible genotypic differences between representatives of many common laboratory mice strains. These genetic differences can be used to select strains which are genetically diverse at an indicated genetic locus and which are thus providing additional information when performing phenotyping or breeding-based mouse experiments. Thus *in silico* fine mapping is a first, very efficient step on the way of unraveling genotype-phenotype relationships.

During the implementation of MouseFM we have paid attention to a very easy handling. To perform a fine mapping study, our tool only requires binary information (e.g. case versus control) for a phenotype of interest on at least two of the 37 available input strains. Further optional parameters can be set to reduce or expand the search space. MouseFM can also be performed on quantitative traits as we showed in the interfrontal bone example.

In conclusion, MouseFM implements a conceptually simple, but powerful approach for *in silico* fine mapping inluding a very comprehensive SNP set of 37 inbred mouse strains. By re-analyzing three fine mapping studies, we demonstrate that MouseFM is a very useful tool for studying genotype-phenotype relationships in mice.

ACKNOWLEDGEMENTS

MM, HB and IW acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under EXC 22167-390884018 and KFO 303 grant number BU 2487/3. All authors acknowledge computational support from the OMICS compute cluster at the University of Lübeck.

REFERENCES

Aherrahrou, Z., Axtner, S. B., Kaczmarek, P. M., Jurat, A., Korff, S., Doehring, L. C., Weichenhan, D., Katus, H. A., and Ivandic, B. T. (2004). A locus on chromosome 7 determines dramatic up-regulation of osteopontin in dystrophic cardiac calcification in mice. *The American Journal of Pathology*, 164(4):1379–1387.

Aherrahrou, Z., Doehring, L. C., Kaczmarek, P. M., Liptau, H., Ehlers, E.-M., Pomarino, A., Wrobel, S., Götz, A., Mayer, B., Erdmann, J., and Schunkert, H. (2007). Ultrafine mapping of Dyscalc1 to an 80-kb chromosomal segment on chromosome 7 in mice susceptible for dystrophic calcification. *Physiological Genomics*, 28(2):203–212.

Ashbrook, D. G., Arends, D., Prins, P., Mulligan, M. K., Roy, S., Williams, E. G., Lutz, C. M., Valenzuela, A., Bohl, C. J., Ingels, J. F., McCarty, M. S., Centeno, A. G., Hager, R., Auwerx, J., Sen, S., Lu, L., and Williams, R. W. (2019). The expanded BXD family of mice: A cohort for experimental systems genetics and precision medicine. *bioRxiv*, page 672097.

Beermann, F., Orlow, S. J., and Lamoreux, M. L. (2004). The Tyr (albino) locus of the laboratory mouse. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 15(10):749–758.

Bogue, M. A., Grubb, S. C., Walton, D. O., Philip, V. M., Kolishovski, G., Stearns, T., Dunn, M. H., Skelly, D. A., Kadakkuzha, B., TeHennepe, G., Kunde-Ramamoorthy, G., and Chesler, E. J. (2018). Mouse Phenome Database: an integrative database and analysis suite for curated empirical phenotype data from laboratory mice. *Nucleic Acids Research*, 46(D1):D843–D850.

Cervino, A. C. L., Darvasi, A., Fallahi, M., Mader, C. C., and Tsinoremas, N. F. (2007). An integrated in silico gene mapping strategy in inbred mice. *Genetics*, 175(1):321–333.

Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., Meehan, T. F., Weninger, W. J., Westerberg, H., Adissu, H., Baker, C. N., Bower, L., Brown, J. M., Caddle, L. B., Chiani, F., Clary, D., Cleak, J., Daly, M. J., Denegre, J. M., Doe, B., Dolan, M. E., Edie, S. M., Fuchs, H., Gailus-Durner, V., Galli, A., Gambadoro, A., Gallegos, J., Guo, S., Horner, N. R., Hsu, C.-W., Johnson, S. J., Kalaga, S., Keith, L. C., Lanoue, L., Lawson, T. N., Lek, M., Mark, M., Marschall, S., Mason, J., McElwee, M. L., Newbigging, S., Nutter, L. M. J., Peterson, K. A., Ramirez-Solis, R., Rowland, D. J., Ryder, E., Samocha, K. E., Seavitt, J. R., Selloum, M., Szoke-Kovacs, Z., Tamura, M., Trainor, A. G., Tudose, I., Wakana, S., Warren, J., Wendling, O., West, D. B., Wong, L., Yoshiki, A., International Mouse Phenotyping Consortium, Jackson Laboratory, Infrastructure Nationale PHENOMIN, Institut Clinique de la Souris (ICS), Charles River Laboratories, MRC Harwell, Toronto Centre for Phenogenomics, Wellcome Trust Sanger Institute, RIKEN BioResource Center, MacArthur, D. G., Tocchini-Valentini, G. P., Gao, X., Flicek, P., Bradley, A., Skarnes, W. C., Justice, M. J., Parkinson, H. E., Moore, M., Wells, S., Braun, R. E., Svenson, K. L., de Angelis, M. H., Herault, Y., Mohun, T., Mallon, A.-M., Henkelman, R. M., Brown, S. D. M., Adams, D. J., Lloyd, K. C. K., McKerlie, C., Beaudet, A. L., Bućan, M., and Murray, S. A. (2016). High-throughput discovery of novel developmental phenotypes. *Nature*, 537(7621):508–514.

Doran, A. G., Wong, K., Flint, J., Adams, D. J., Hunter, K. W., and Keane, T. M. (2016). Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biology*, 17(1):167.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44.

Grupe, A., Germer, S., Usuka, J., Aud, D., Belknap, J. K., Klein, R. F., Ahluwalia, M. K., Higuchi, R., and Peltz, G. (2001). In silico mapping of complex disease-related traits in mice. *Science (New York, N.Y.)*, 292(5523):1915–1918.

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I. M., Trevanion, S. J., Flicek, P., and Cunningham, F. (2018). Ensembl variation resources. *Database*, 2018. bay119.

Ivandic, B. T., Qiao, J. H., Machleder, D., Liao, F., Drake, T. A., and Lusis, A. J. (1996). A locus on chromosome 7 determines myocardial cell necrosis and calcification (dystrophic cardiac calcinosis) in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 93(11):5483–5488.

Ivandic, B. T., Utz, H. F., Kaczmarek, P. M., Aherrahrou, Z., Axtner, S. B., Klepsch, C., Lusis, A. J., and Katus, H. A. (2001). New Dyscalc loci for myocardial cell necrosis and calcification (dystrophic cardiac calcinosis) in mice. *Physiological Genomics*, 6(3):137–144.

Jeong, E., Choi, H. K., Park, J. H., and Lee, S. Y. (2018). STAC2 negatively regulates osteoclast formation by targeting the RANK signaling complex. *Cell Death and Differentiation*, 25(8):1364–1374.

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., McIntyre, R. E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assunção, J. A., Donahue, L. R., Reinholdt, L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J., and Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294.

Kim, H.-G., Kim, H.-T., Leach, N. T., Lan, F., Ullmann, R., Silahtaroglu, A., Kurth, I., Nowka, A., Seong, I. S., Shen, Y., Talkowski, M. E., Ruderfer, D., Lee, J.-H., Glotzbach, C., Ha, K., Kjaergaard, S., Levin, A. V., Romeike, B. F., Kleefstra, T., Bartsch, O., Elsea, S. H., Jabs, E. W., MacDonald, M. E., Harris, D. J., Quade, B. J., Ropers, H.-H., Shaffer, L. G., Kutsche, K., Layman, L. C., Tommerup, N., Kalscheuer, V. M., Shi, Y., Morton, C. C., Kim, C.-H., and Gusella, J. F. (2012). Translocations disrupting PHF21A in the Potocki-Shaffer-syndrome region are associated with intellectual disability and craniofacial anomalies. *American Journal of Human Genetics*, 91(1):56–72.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122.

Meehan, T. F., Conte, N., West, D. B., Jacobsen, J. O., Mason, J., Warren, J., Chen, C.-K., Tudose, I., Relac, M., Matthews, P., Karp, N., Santos, L., Fiegel, T., Ring, N., Westerberg, H., Greenaway, S., Sneddon, D., Morgan, H., Codner, G. F., Stewart, M. E., Brown, J., Horner, N., International Mouse Phenotyping Consortium, Haendel, M., Washington, N., Mungall, C. J., Reynolds, C. L., Gallegos, J., Gailus-Durner, V., Sorg, T., Pavlovic, G., Bower, L. R., Moore, M., Morse, I., Gao, X., Tocchini-Valentini, G. P., Obata, Y., Cho, S. Y., Seong, J. K., Seavitt, J., Beaudet, A. L., Dickinson, M. E., Herault, Y., Wurst, W., de Angelis, M. H., Lloyd, K. C. K., Flenniken, A. M., Nutter, L. M. J., Newbigging, S., McKerlie, C., Justice, M. J., Murray, S. A., Svenson, K. L., Braun, R. E., White, J. K., Bradley, A., Flicek, P., Wells, S., Skarnes, W. C., Adams, D. J., Parkinson, H., Mallon, A.-M., Brown, S. D. M., and Smedley, D. (2017). Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nature Genetics*, 49(8):1231–1238.

Meng, H., Vera, I., Che, N., Wang, X., Wang, S. S., Ingram-Drake, L., Schadt, E. E., Drake, T. A., and Lusis, A. J. (2007). Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proceedings of the National Academy of Sciences of the United States of America*, 104(11):4530–4535.

Mulligan, M. K., Mozhui, K., Prins, P., and Williams, R. W. (2017). GeneNetwork: A Toolbox for Systems Genetics. *Methods in Molecular Biology (Clifton, N.J.)*, 1488:75–120.

Uhl, E. W. and Warner, N. J. (2015). Mouse Models as Predictors of Human Responses: Evolutionary Medicine. *Current Pathobiology Reports*, 3(3):219–223.

Wade, C. M., Kulbokas, E. J., Kirby, A. W., Zody, M. C., Mullikin, J. C., Lander, E. S., Lindblad-Toh, K., and Daly, M. J. (2002). The mosaic structure of variation in the laboratory mouse genome. *Nature*, 420(6915):574–578.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. S., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P. (2014). The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*, 31(1):143–145.

Zimmerman, H., Yin, Z., Zou, F., and Everett, E. T. (2019). Interfrontal Bone Among Inbred Strains of Mice and QTL Mapping. *Frontiers in Genetics*, 10:291.