Published online 25 October 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D753–D756***
doi: 10.1093/nar/gkv1106

NAR Breakthrough Article

dbMAE: the database of autosomal monoallelic expression

Virginia Savova[1,2,*], Jon Patsenker[1], Sébastien Vigneau[1] and Alexander A. Gimelbrant[1,*]

[1]Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, 450 Brookline Ave., Boston, MA 02215, USA and [2]Department of Systems Biology, Harvard Medical School, 200 Longwood Ave., Boston, MA 02215, USA

Received August 18, 2015; Revised October 05, 2015; Accepted October 11, 2015

ABSTRACT

Recently, data on 'random' autosomal monoallelic expression has become available for the entire genome in multiple human and mouse tissues and cell types, creating a need for better access and dissemination. The database of autosomal monoallelic expression (dbMAE; https://mae.hms.harvard.edu) incorporates data from multiple recent reports of genome-wide analyses. These include transcriptome-wide analyses of allelic imbalance in clonal cell populations based on sequence polymorphisms, as well as indirect identification, based on a specific chromatin signature present in MAE gene bodies. Currently, dbMAE contains transcriptome-wide chromatin identification calls for 8 human and 21 mouse tissues, and describes over 16 000 murine and ∼700 human cases of directly measured biased expression, compiled from allele-specific RNA-seq and genotyping array data. All data are manually curated. To ensure cross-publication uniformity, we performed re-analysis of transcriptome-wide RNA-seq data using the same pipeline. Data are accessed through an interface that allows for basic and advanced searches; all source references, including raw data, are clearly described and hyperlinked. This ensures the utility of the resource as an initial screening tool for those interested in investigating the role of monoallelic expression in their specific genes and tissues of interest.

INTRODUCTION

Autosomal monoallelic expression (MAE) refers to mitotically stable, epigenetically controlled allele-specific expression of autosomal genes, with the initial non-predetermined ('random') choice of the transcriptional activity of the two alleles maintained in a given clonal cell lineage [recent reviews include (1–3)]. It is thought to result in massive diversity between cells within the same tissue.

MAE is a relatively recent addition to the number of known epigenetic mechanisms that involve separate regulation of the two alleles' activity in mammalian cells, including X-chromosome inactivation (4), allelic exclusion in immunoglobulin loci (5) and genomic imprinting (6). After discovery of random allelic inactivation of olfactory receptor genes (7), there were scattered reports of MAE in a variety of mouse genes, including *Il4* (8) and innate immunity receptor *Tlr4* (9).

The advent of transcriptome-wide approaches – based first on hybridization to arrays (10,11) or beads (12) and then RNA sequencing (13–16) – revealed that MAE is widespread in human and mouse cells. Analysis of allelic expression in a limited number of clonal cell lines indicated that a small percentage of genes in a given cell type were subject to MAE. Importantly, for most genes subject to MAE, in some clonal lines the gene was expressed equally from both alleles. This makes the detection of MAE strongly dependent on the number of clonal lines available: the more lines analyzed, the higher the likelihood that MAE status for a gene will be detected in at least one line.

In addition to isogenic clonal cell lines, the approaches used in these studies depend on the existence of polymorphisms between maternal and paternal copies of a gene. Accordingly, much higher polymorphism density in F1 hybrid mouse crosses [e.g. (11)] compared to heterozygosity in hu-

*To whom correspondence should be addressed. Tel: +1 617 582 7377; Fax: +1 617 632 4770; Email: Virginia_Savova@hms.harvard.edu
Correspondence may also be addressed to Alexander A. Gimelbrant. Tel: +1 617 582 7326; Fax: +1 617 632 4770; Email: gimelbrant@mail.dfci.harvard.edu

man samples [e.g. (10)] resulted in a much greater fraction of genes being informative.

Recently, it was shown that MAE can be identified regardless of sequence variation, on the basis of a characteristic signature of chromatin modifications in human (13) and mouse (14) cells. Based on this approach, the ability to classify genes as MAE or biallelically expressed has become available for the entire genome in multiple tissues and cell types in two organisms, human and mouse, creating a need for better data access and dissemination.

THE MAE DATABASE

The database of autosomal MAE genes described here is targeted to researchers interested in investigating MAE in individual genes or groups of genes. We compiled information from multiple reports of allele-specific expression and chromatin signature of MAE genes. Note that the relationship between chromatin modifications we assessed and allele-specific silencing has not been established for imprinted genes (17) or genes on the X chromosome. The same is the case of olfactory receptor genes, the largest gene family in mammalian genomes, for which expression of one allele per neuron is thought to be obligatory (18). Finally, we did not incorporate data from single-cell sequencing studies of allele-specific expression [e.g. (19)] since it is not yet clear how to distinguish mitotically stable autosomal MAE from stochastic transcription in single cells.

Data organization, accuracy and uniformity

dbMAE contains two broad classes of data: direct measurement of allelic expression (im)balance (termed 'experimental') and indirect chromatin-based inference ('inferred'). Only data from peer-reviewed publications is included. The database maintains a clear distinction between these types of evidence, listing the entry's MAE status according to all available sources. The user can also access the full set of data for each gene.

The data are organized into separate tables containing gene information in the form of unique database ID – a gene symbol, or RefSeq/Entrez ID; and experimental or prediction data for each unique gene ID in each tissue. At the time of this publication, transcript-specific data on monoallelic expression is not available. For example, (13,14) considered only the longest transcript for that gene, without parsing reads between overlapping isoforms. However, the database structure allows for seamless integration of future transcript specific data.

The allelic expression data (RNA-seq and microarray) is taken from six recent studies (10,11,13–15). To ensure uniform criteria of monoallelic and biallelic expression, the data are analyzed and curated according to the method described in (13). Briefly, biased expression in a clonal sample is called if a gene's cumulative variant counts on phased SNPs pass the FDR-corrected binomial test, and the gene is biased at least 2:1 in favor of either allele. The former condition ensures that technical noise due to low coverage will not be translated into monoallelic calls. The latter condition buffers against the binomial test's hypersensitivity to small deviations from 50:50 in highly covered genes. Further, empirical monoallelic expression in a cell type is called if at least one clonal line belonging to this cell type has shown monoallelic expression. Note that some instances of biased expression may be due to genetic bias rather than epigenetic MAE. Thus the experimental data are best interpreted in the broader context of all data available for the gene, and in conjunction with the predictions from chromatin analysis.

The inference of MAE from chromatin signature, described and validated in (13,14), involves a decision-tree utilizing information about the overall input-normalized enrichment of H3K27me3 and H3K36me3 integrated over the gene body. The co-occurrence of those marks is a strong indication of MAE. Inferred monoallelic or biallelic expression status in a given tissue is assigned if the gene: (i) has the chromatin signature of MAE in that tissue and (ii) is ranked on the top half of all genes by expression level. This standard cut-off was adopted in (14) and retroactively applied to data from (13) to achieve a unified approach. Since MAE is known to be a tissue-specific phenomenon, it is sufficient for a gene to be inferred MAE in one tissue to be classified as 'Monoallelic' in the database. To achieve inferred 'Biallelic' status, a gene must not be inferred monoallelic in any tissue. In addition to the categories 'Monoallelic' and 'Biallelic', we include a category 'Undetermined', which indicates that evidence for MAE exists only from chromatin-based method and is detected only at very low expression levels, where its validity has not been extensively studied.

At the time of publication, the database contains MAE calls for over 700 human and 16 000 mouse genes with experimental evidence of allelic bias as assessed in four tissue types. There are close to 5000 corresponding positively confirmed instances of biallelic expression in human and more than 21 000 in mouse. In addition, the database includes MAE and biallelic calls based on genome-wide chromatin analysis for seven human and 21 murine cell types and tissues.

Usability and maintenance

The database is publicly available and searchable via http://mae.hms.harvard.edu. Searchable fields include gene name, cell or tissue type and organism as well as gene allelic expression status. The gene name field accepts common gene names, Ensembl, RefSeq and Entrez IDs.

Search functionality is divided into basic and advanced modes (see online supplementary). In basic mode, the user can check if a gene is inferred or measured to have monoallelic expression in mouse or human tissues (e.g. *Msx2*), or how classes of genes behave (by checking the 'Search for gene class prefix' option and entering a partial name, e.g. 'Msx').

More complex searches can be carried out after clicking the 'Advanced' button. For example, it is possible to check the status of a gene in a specific organism and tissue (e.g. 'Msx2' with Organism: 'Mouse' and Tissue: 'NeuronalProgenitor_GSE54016'). Advanced mode also allows searching by status to check if a class of genes is inferred/known to be monoallelically expressed anywhere (e.g. 'Msx2' with Status: 'Monoallelic' and Organism/Tissue: 'Any organism'), or if a gene is inferred/known to show monoallelic expression in mouse fibroblasts specifically ('Msx2' with Status: 'Monoallelic' and Organism/Tissue: 'Mouse Fibroblast').

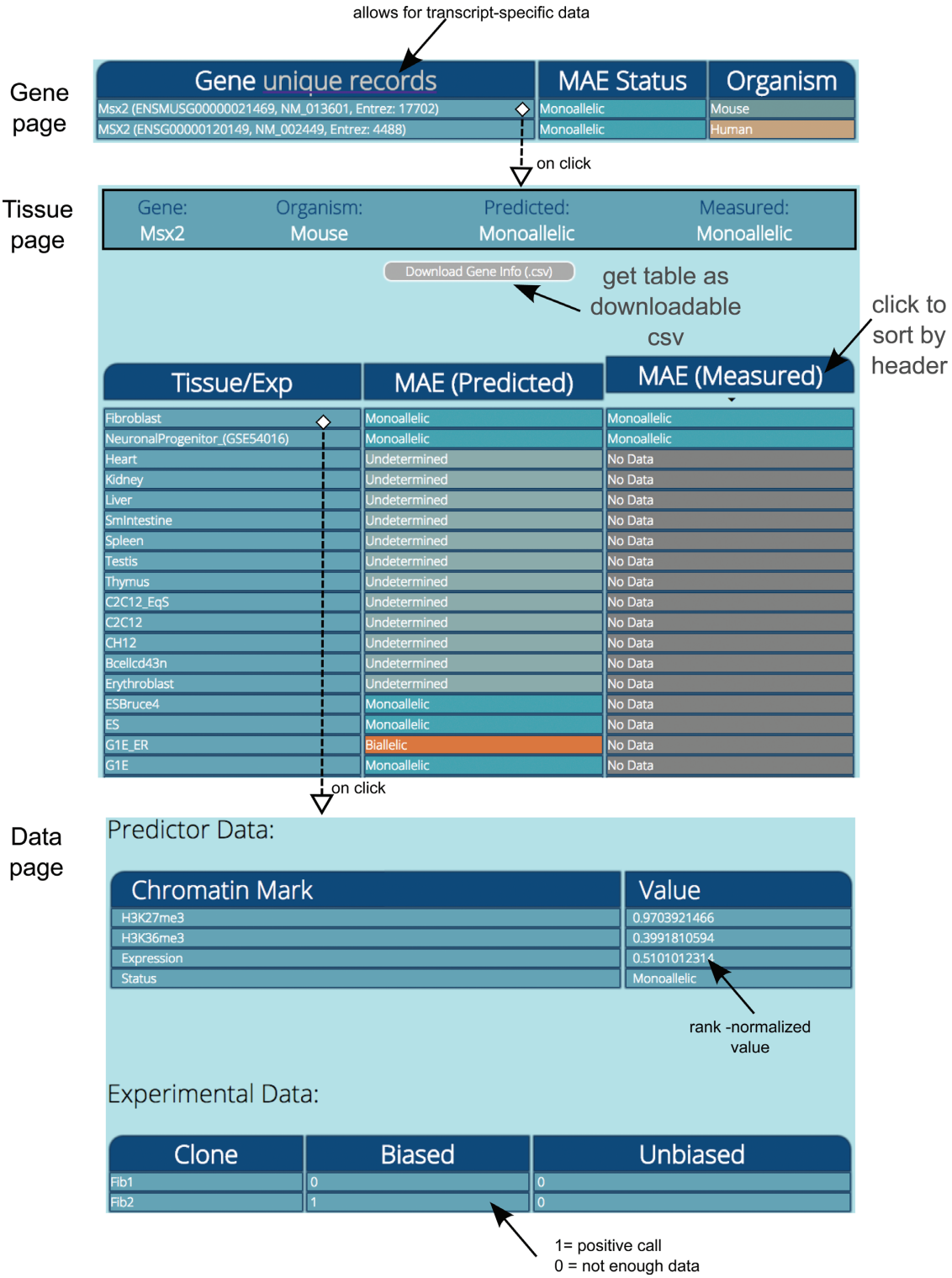

Figure 1. Results of queries to the databases can be examined to greater specificity by clicking on fields of interest. Shown are data panels from pages resulting from simple query by gene symbol 'Msx2'. ***Gene* page:** Indicates overall gene status in the organisms currently in the database. 'Monoallelic' indicates that there is some evidence for monoallelic expression in at least one tissue. ***Tissue* page:** Accessed upon clicking one of the entries in the 'gene page'; indicates inferred and experimental status of the gene in the selected organism in each tissue. Headers can be clicked to sort alphabetically. Clicking the 'Download as csv' button results in a csv file with the data as presented on the page. ***Data* page:** Accessed upon clicking one of the entries in the 'tissue page'. The chromatin mark section shows quantile rank values for the chromatin marks H3K27me3 and H3K36me3 in the selected tissue for the query gene. The experimental data section shows clonal cell lines derived from the tissue in question, and whether the query gene was positively called ("1") or undetermined ("0") to be either *biased* or *unbiased* in each line.

To maintain a balance between providing the user with all available evidence and presenting it in a usable form, as a first step, we indicate whether a gene shows any evidence for MAE for a given context specified during the search (e.g. in any tissue type in either organism, or in a specific tissue/organism). Thus, a user is first presented with all matching searches in a summary page, where each matching entry is described as 'monoallelic' if it is either inferred or empirically determined to be monoallelic in at least one tissue (Figure 1). Clicking on the entry of interest opens the tissue page, where inferred and experimental data are described for each tissue or cell type. Clicking on the name of a specific tissue shows data used for prediction and/or allelic expression data from individual clones.

In addition, we have implemented a basic batch mode, where multiple comma-separated gene IDs can be entered. This yields all individual database entries for these IDs, in a comma-separated format.

Note that the inferred and experimental data do not necessarily come from the same biological sample, but are nevertheless matched for comparison purposes. For example, the experimental data on neuronal progenitor cells derived from datasets GSE54016 and E-MTAB-1822 are presented as separate datasets, since they involve different experiments from different labs. At the same time, both are compared with the same chromatin-based inference on neuronal progenitor samples.

Furthermore, measured and inferred allelic expression statuses are not always identical for the same cell or tissue type. Biological differences between samples, for instance due to genetic background or clonal derivation, may explain some of these inconsistencies. Limitations intrinsic to both approaches may also play a role. For instance, when assessing MAE using RNA-Seq in heterozygous clonal lines, the number of clonal lines assessed may not be sufficient to capture MAE that normally happens in a small proportion of cells. Conversely, chromatin signature may capture unstable MAE, or be confounded by bimodal distribution of expression across cells (e.g. if a gene is completely shut down in some cells, but is biallelic in other cells, we will observe a co-occurrence of chromatin marks for silencing and active transcription similar to those observed with MAE) [see (13,14)]. These approaches are complementary and presenting allelic expression and chromatin signature data side by side should be beneficial to users.

The database is maintained by the corresponding authors. Submissions of new or updated information based on peer-reviewed publications are encouraged.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank early users for their comments and suggestions: Jason Alvarez, Suzanne Gaudet, Marko Horb, Anwesha Nag, Clara Pereira and Leon Peshkin. We would also like to thank the anonymous reviewers for the improvements that resulted from their testing, and Andy Bergman, an Orchestra system administrator, for his help and support. dbMAE is hosted on the Orchestra High Performance Compute Cluster at Harvard Medical School. See http://rc.hms.harvard.edu for more information.

FUNDING

Pew scholar award [to A.A.G. in part]; NIH [R01GM114864 to V.S., S.V., A.A.G in part]. This NIH supported shared facility is partially provided through NCRR 1S10RR028832–01. Funding for open access charge: NIH.
Conflict of interest statement. None declared.

REFERENCES

1. Chess,A. (2012) Mechanisms and consequences of widespread random monoallelic expression. *Nat. Rev. Genet.* , **13**, 421–428.
2. Savova,V., Vigneau,S. and Gimelbrant,A.A. (2013) Autosomal monoallelic expression: genetics of epigenetic diversity? *Curr. Opin. Genet. Dev.*, **23**, 642–648.
3. Eckersley-Maslin,M.A. and Spector,D.L. (2014) Random monoallelic expression: regulating gene expression one allele at a time. *Trends Genet.*, **30**, 237–244.
4. Lee,J.T. and Bartolomei,M.S. (2013) X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell*, **152**, 1308–1323.
5. Levin-Klein,R. and Bergman,Y. (2014) Epigenetic regulation of monoallelic rearrangement (allelic exclusion) of antigen receptor genes. *Frontiers Immunol.*, **5**, 625.
6. Plasschaert,R.N. and Bartolomei,M.S. (2014) Genomic imprinting in development, growth, behavior and stem cells. *Development*, **141**, 1805–1813.
7. Chess,A., Simon,I., Cedar,H. and Axel,R. (1994) Allelic inactivation regulates olfactory receptor gene expression. *Cell*, **78**, 823–834.
8. Bix,M. and Locksley,R.M. (1998) Independent and epigenetic regulation of the interleukin-4 alleles in CD4+ T cells. *Science (New York, N.Y.)*, **281**, 1352–1354.
9. Pereira,J.P., Girard,R., Chaby,R., Cumano,A. and Vieira,P. (2003) Monoallelic expression of the murine gene encoding Toll-like receptor 4. *Nat. Immunol*, **4**, 464–470.
10. Gimelbrant,A., Hutchinson,J.N., Thompson,B.R. and Chess,A. (2007) Widespread monoallelic expression on human autosomes. *Science (New York, N.Y.)*, **318**, 1136–1140.
11. Zwemer,L.M., Zak,A., Thompson,B.R., Kirby,A., Daly,M.J., Chess,A. and Gimelbrant,A.A. (2012) Autosomal monoallelic expression in the mouse. *Genome Biol.*, **13**, R10.
12. Jeffries,A.R., Perfect,L.W., Ledderose,J., Schalkwyk,L.C., Bray,N.J., Mill,J. and Price,J. (2012) Stochastic choice of allelic expression in human neural stem cells. *Stem Cells*, **30**, 1938–1947.
13. Nag,A., Savova,V., Fung,H.L., Miron,A., Yuan,G.C., Zhang,K. and Gimelbrant,A.A. (2013) Chromatin signature of widespread monoallelic expression. *Elife*, **2**, e01256.
14. Nag,A., Vigneau,S., Savova,V., Zwemer,L.M. and Gimelbrant,A.A. (2015) Chromatin signature identifies monoallelic gene expression across mammalian cell types. *G3*, **5**, 1713–1720.
15. Gendrel,A.V., Attia,M., Chen,C.J., Diabangouaya,P., Servant,N., Barillot,E. and Heard,E. (2014) Developmental dynamics and disease potential of random monoallelic gene expression. *Dev. Cell*, **28**, 366–380.
16. Eckersley-Maslin,M.A., Thybert,D., Bergmann,J.H., Marioni,J.C., Flicek,P. and Spector,D.L. (2014) Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev. Cell*, **28**, 351–365.
17. Morison,I.M., Ramsay,J.P. and Spencer,H.G. (2005) A census of mammalian imprinting. *Trends Genet.*, **21**, 457–465.
18. Magklara,A. and Lomvardas,S. (2013) Stochastic gene expression in mammals: lessons from olfaction. *Trends Cell Biol.* , **23**, 449–456.
19. Deng,Q., Ramskold,D., Reinius,B. and Sandberg,R. (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science (New York, N.Y.)*, **343**, 193–196.