


METHODOLOGY

Open Access



Medical Subject Heading (MeSH) annotations illuminate maize genetics and evolution

Timothy M. Beissinger^{1*}  and Gota Morota²

Abstract

Background: High-density marker panels and/or whole-genome sequencing, coupled with advanced phenotyping pipelines and sophisticated statistical methods, have dramatically increased our ability to generate lists of candidate genes or regions that are putatively associated with phenotypes or processes of interest. However, the speed with which we can validate genes, or even make reasonable biological interpretations about the principles underlying them, has not kept pace. A promising approach that runs parallel to explicitly validating individual genes is analyzing a set of genes together and assessing the biological similarities among them. This is often achieved via gene ontology analysis, a powerful tool that involves evaluating publicly available gene annotations. However, additional resources such as Medical Subject Headings (MeSH) can also be used to evaluate sets of genes to make biological interpretations.

Results: In this manuscript, we describe utilizing MeSH terms to make biological interpretations in maize. MeSH terms are assigned to PubMed-indexed manuscripts by the National Library of Medicine, and can be directly mapped to genes to develop gene annotations. Once mapped, these terms can be evaluated for enrichment in sets of genes or similarity between gene sets to provide biological insights. Here, we implement MeSH analyses in five maize data-sets to demonstrate how MeSH can be leveraged by the maize and broader crop-genomics community.

Conclusions: We demonstrate that MeSH terms can be effectively leveraged to generate hypotheses and make biological interpretations in maize, and we provide a pipeline that enables the use of MeSH terms in other plant species.

Keywords: MeSH, Maize, Gene ontology (GO), Overrepresentation analysis (ORA), Domestication, Ear number, Seed size, Inflorescence

Background

Technological advances in sequencing and phenotyping have accelerated in recent decades, enabling high-throughput studies aimed at associating genotypes and phenotypes. In many cases, the speed at which we can generate large sets of candidate associations from genome-wide association studies (GWAS) [1], selection mapping [2], and other approaches has surpassed our ability to draw meaningful biological conclusions from

these candidates. However, as was recently described by Rausher and Delph [3], gene-identification is not always necessary to draw meaningful insights. Alternatively, it is often possible to look for recurrent patterns among distinct sets of candidate genes or regions in order to elucidate meaning. Annotation-based tests for enrichment or similarity represent one avenue for unraveling meaning from sets of candidates. In brief, these approaches involve identifying statistically enriched annotation terms among a list of candidate sites (usually genes or regions), or looking for similarity between terms corresponding to two sets of candidate sites, and inferring that there may be a biological explanation for the enriched or similar terms.

*Correspondence: Tim.beissinger@ars.usda.gov

¹ USDA-ARS Plant Genetics Research Unit, Division of Plant Sciences, Division of Biological Sciences, MU Informatics Institute, University of Missouri, Columbia, MO 65211, USA

Full list of author information is available at the end of the article



Commonly applied techniques often utilize gene ontology (GO) annotations [4], which provide putative descriptions of gene function [5, 6]. GO annotations are an important genomic tool to provide insight into biological interpretations of gene sets. However, despite their well-proven utility, there is growing interest in additional annotation-based approaches that can be leveraged to complement, support, enhance, or add to the patterns identified by GO. Included among this assortment of strategies are KEGG annotations [7], Disease Ontology [8], and Medical Subject Headings (MeSH), which were introduced at the National Library of Medicine (NLM) more than 50 years ago [9].

MeSH terms are the NLM's controlled terminology, primarily used to organize and index information and manuscripts found in common databases such as PubMed [10]. By mapping from MeSH terms to manuscripts, and then to a list of candidate genes, a semantic pattern search for biological meaning can be conducted [11]. Recently, the MeSH Over-representation Analysis (ORA) Framework, a suite of software for conducting MeSH enrichment analyses using R [12] and Bioconductor [13], was developed [14]. MeSH analysis has proven useful for deducing meaning from sets of genes implicated across several agricultural animal species including in cattle, swine, horse and chicken [15, 16]. Here, we implement five MeSH analyses in maize, which collectively demonstrate how MeSH can be used to enrich biological understanding in crop species.

In this study, which is meant to be both a primer for MeSH-based analysis in maize and other crop plants, as well as an investigation of patterns that can be deduced regarding maize genetics and evolution, we identify over-represented MeSH terms among candidate genes identified from five distinct maize datasets: (1) regions under selection during maize domestication [17]; (2) regions under selection during maize improvement [17]; (3) regions under selection for seed size [18]; (4) regions under selection for ear number [19]; and (5) regions contributing to inflorescence traits [20]. After identifying significant MeSH terms, we also assess and test for semantic similarity, or MeSH-based relatedness, among

the genes identified in each of these datasets to identify relationships among the genetic underpinnings of these traits/selection regimes.

Methods

Code availability

To enable implementation of MeSH analyses by other researchers, all scripts used in this study are available as annotated additional files in R-markdown format (Additional files 1, 2, 3, 4, 5, 6, 7). Scripts were written in R [12] and utilize Bioconductor [13], the MeSH ORA Framework including the “meshr” for ORA and the “Mesh.Zma.e.g.db” maize-specific mapping table [14], and MeSHSim [21]. The mapping table provides the necessary link between NCBI Entrez Gene IDs and NLM MeSH IDs. For maize, the mapping table was provided by gene2pubmed [22] with data licensed by PubMed. The GOstats R package [23] was used to implement GO ORA to generate a baseline that MeSH results could be compared to. Genome data was downloaded using the biomaRt R package [24]. Full analysis details are included within the reproducible scripts (Additional files 1, 2, 3, 4, 5, 6, 7).

Datasets

We analyzed five publicly available datasets to identify enriched MeSH terms and look for semantic similarity between different traits and selection regimes. The datasets analyzed are described in Table 1. For the four datasets that involved contiguous regions (Domestication, improvement, seed size, and ear number), all genes that fell within the implicated regions were used for MeSH analysis. For the remaining dataset (inflorescence traits), which involved isolated SNPs identified through GWAS instead of genomic regions, all genes within 10 kb of the implicated SNPs were used for MeSH analysis. All gene models and gene locations were based on the maize reference genome version 2 [25].

Analyses

Each of the five datasets was first tested for any over-represented MeSH terms and GO terms. MeSH ORA

Table 1 Datasets used in this study, including reference information where full details can be found and a brief description of each

Dataset	Reference	Description
Domestication	Hufford et al. [17]	Regions selected during domestication from teosinte to maize
Improvement	Hufford et al. [17]	Regions selected during post-domestication maize improvement
Seed size	Hirsch et al. [18]	Regions artificially selected for seed size in a long-term selection experiment
Ear number	Beissinger et al. [19]	Regions artificially selected for ear number in a long-term selection experiment
Inflorescence traits	Brown et al. [20]	SNPs associated with inflorescence traits from a genome-wide association study

was performed using the MeSH ORA Framework which includes the “meshr” and “MeSH.Zma.e.g.db” R-packages [14], the latter of which is a mapping table that connects gene Entrez Gene IDs to MeSH IDs. These packages can be installed using Bioconductor by running the command, “source(“https://bioconductor.org/biocLite.R”); followed by “biocLite(“meshr”)” and “biocLite(“MeSH.Zma.e.g.db”)”. Further instructions to install and run these packages are provided in Additional files 1, 2, 3, 4 and 5. Unfortunately, the majority of maize genes annotated in the maize version 2 reference genome [25] do not have a corresponding Entrez Gene ID, and therefore are not useful for MeSH analyses. Of the 40,481 gene models available from Ensembl Plants [26], only 14,142 have corresponding Entrez IDs. The “meshHyperGTest” function was implemented to conduct a hypergeometric test. Specifically, to test the probability that a specific MeSH term is enriched in a particular set of genes, as compared to a background gene set, this function calculates

$$P(\text{enrichment}) = \sum_{x=s}^{\min(M,k)} \frac{\binom{M}{x} \binom{N-M}{k-x}}{\binom{N}{k}},$$

where N is the total number of background genes, k is the number of genes in the set being tested, M is the number of background genes corresponding to the particular MeSH term, and s is the number of genes in the test set that correspond to that MeSH term [14]. For this study, all Entrez genes in the maize reference genome version 2 [25] were used as the background gene set. GO ORA was conducted using a similar approach, as demonstrated in the additional files. The necessary GOSTats package, which requires a list of Entrez Gene IDs as input, is installed by running “biocLite(“GOSTats”)”.

Next, semantic similarity between distinct experiments was evaluated using the MeSHSim R package [21] to elucidate if there are underlying relationships between the

trait data-sets (seed size, ear number, or inflorescence traits) and the process data-sets (domestication, improvement), as well as the relationships within the process and trait datasets. The “headingSetSim” function was used, and results were plotted with the corrplot R package [27].

Results

Overrepresentation analysis

MeSH ORA involves performing a hypergeometric test to determine which MeSH terms are enriched among the candidate set of genes compared to a set of background genes. All genes in the maize reference genome version 2 [25] with Entrez Gene IDs were used as the background set. While GO terms are classified into the three groups “molecular function”, “cellular components”, and “biological processes”, MeSH classifications include several groups, many of which are geared more toward indexing biomedical manuscripts than biological processes. However, classifications including “chemicals and drugs”, “diseases”, “anatomy”, and “phenomena and processes”, all have the potential to contribute to the biological understanding of sets of genes. Counts of the number of overrepresented terms in three classification groups for MeSH and GO are provided in Table 2. The precise overrepresented terms in each of these categories for the five analyzed datasets are described in Additional files 1, 2, 3, 4 and 5. For the purpose of demonstration, MeSH terms identified within the “anatomy” classification are provided as an example and described in detail in Table 3. Many of the enriched terms serve to provide additional evidence for reasonable a priori expectations, such as the observation that “flowers” and “seeds” are both enriched within the set of genes under selection during domestication. However, others introduce interesting questions that could serve to drive hypothesis generation for future studies. For instance, the only enriched term identified from the ear number dataset is “endosperm”, which one would not immediately assume to be related to ear number.

Table 2 Number of MeSH and GO terms identified within three classification groups for both MeSH and GO

	Domestication	Improvement	Seed size	Ear number	Inflorescence traits
<i>MeSH category</i>					
Chemicals and drugs	18	19	11	0	13
Anatomy	5	7	3	1	4
Phenomena and processes	30	8	18	1	11
<i>GO category</i>					
Biological processes	52	48	59	28	72
Molecular function	27	37	20	17	33
Cellular components	12	15	14	6	8

Table 3 MeSH terms enriched in each of the five datasets within the “anatomy” MeSH classification group

Domestication	Improvement	Seed size	Ear number	Inflorescence traits
<i>MeSH terms</i>				
Chromosomes	Xylem	Cytosol	Endosperm	Endo. reticulum
Centromere	Phloem	Shoots		Cell membrane
Flowers	Chromosomes	Chromosomes		Plant leaves
Seeds	golgi Apparatus			Thylakoids
Cyto. vesicles	Cyto. vesicles			
	Ribosomes			
	Flowers			

Semantic similarity analysis

Another powerful use of MeSH is that it can be used to calculate the semantic similarity between distinct sets of MeSH terms. This type of analysis enables one to look for hidden relationships among sets of genes, potentially uncovering biological meaning. For the five datasets we studied, we assessed whether there were pairwise relationships linking any of them. Figure 1 depicts the MeSH similarity between each set of candidate genes. Interestingly, the strongest relationship identified was between domestication genes and seed size genes, possibly suggesting that seed size traits were more strongly selected during domestication than were ear number or other inflorescence traits. Noteworthy relationships were also observed between domestication and improvement genes, as well as between seed size and improvement genes. It should be noted that ear number genes were not strongly related to any of the other gene sets, which may simply result from the fact that the ear number dataset included the fewest candidate genes. This possibility is elaborated upon further in the discussion.

Comparison of real data to a random set of genes

We conducted an analysis of 1500 randomly selected genes to determine the robustness of MeSH analyses in a scenario where no biological meaning is present (Additional file 6). As is expected for any *p* value based method, a subset of terms achieved significance. Spurious results were also observed in a parallel GO analysis (Additional file 6). In contrast to many of the real datasets we evaluated, there was no overwhelming theme tying the terms together. This subjective observation is supported by a semantic similarity analysis between the random gene set and the real datasets, where lower similarities were generally observed (Additional file 7). Still, the observation that “significant” MeSH or GO terms can arise from a random set of genes suggests that caution should be exercised when attempting to make interpretations from any such study, as is discussed in detail by Pavlidis et al. [28]. Although we utilized a lenient $p = 0.05$ significance threshold here, in part for the purpose of demonstration, the use of a hypergeometric

distribution for testing allows a more stringent significance threshold to be employed when needed.

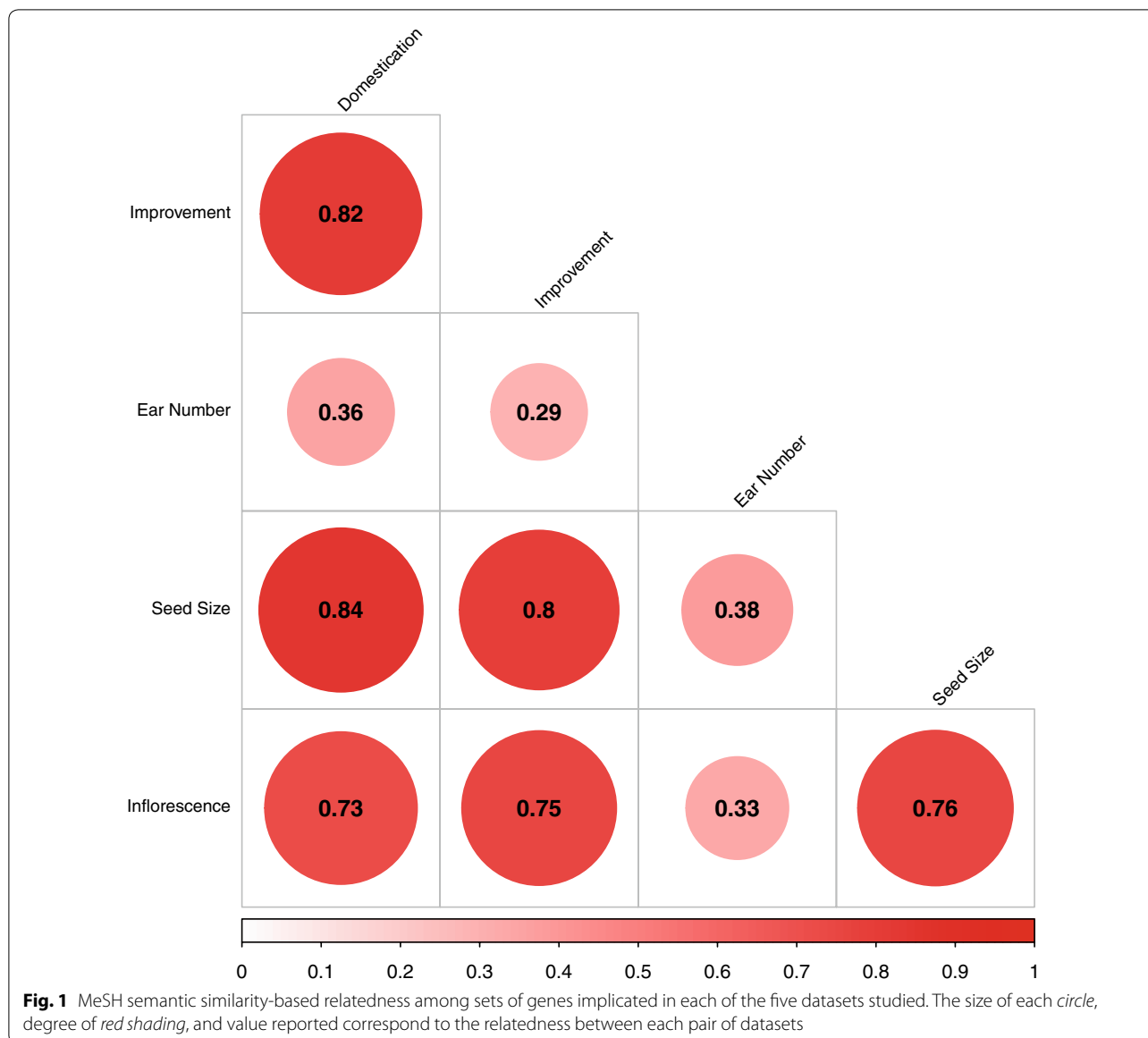
Discussion

Our analysis of five existing datasets demonstrates how MeSH ORA and semantic-similarity analyses can be used to mine data and confirm and/or generate informative hypotheses. Like GO, MeSH-based approaches leverage curated annotations to provide biological insights. In fact, as we have shown, several of the enriched terms within the “anatomy” category are directly related to macro phenotypes, such as “seeds”, “shoots”, “flowers”, and “ears”. Whether applied to existing data, as we have demonstrated here, or if used to infer meaning from a list of candidates generated from a novel mapping study, MeSH represents an additional tool for drawing inferences from large-scale sets of genomic data.

Biological implications

Among the findings gleaned from this analysis was the observation that while both “flowers” and “seeds” were enriched terms in the domestication set of genes, only “flowers” remained significant among improvement genes (Table 3). This result is consistent with the morphological observation that the maize female inflorescence is dramatically different from that of teosinte [29], with one of the most immediately apparent differences being seed related; the teosinte outer glume forms a hard teosinte fruitcase that completely encapsulates each kernel, while in maize the outer glume is barely present [30]. It has been shown that this trait is controlled by relatively few genes, with *tga1* [31, 32] being of particular importance, and therefore our MeSH finding may suggest that after intense selection on seed traits during domestication, subsequent selection on further seed modifications during improvement has possibly been more subdued.

The hypothesis that domestication immediately impacted seed-related traits more than others is further supported by our semantic similarity analysis, where the most similar pair of gene-sets we tested corresponded to domestication and seed size (Fig. 1). Also, while the



limited number of genes included in the ear-number dataset [19] seems to constrain the estimated similarity between ear-number genes and the other datasets, we do observe that ear-number genes are semantically more similar to domestication genes than they are to improvement genes (Fig. 1). This again is consistent with morphological differences between maize and teosinte, with maize demonstrating apical dominance while teosinte has a much more branched structure [33]. The observation of greater similarity between ear number genes and domestication genes than between ear number genes and improvement genes lends support to the existing supposition that single-eared plants have likely been favorable throughout the era of post-domestication maize improvement due to the ease with which such plants can be hand harvested [34].

An observation that ran contrary to our expectation was that “shoots” was an enriched term among seed size genes, while “endosperm” was enriched within the set of ear number genes (Table 3). We are tempted to dismiss these findings as spurious, but both have plausible biological explanations. In the Krug selection population [18], where our seed size regions were identified, mass selection not only impacted seed size, but also affected seedling size, leaf width, stalk circumference, and cob weight [35], indicating that the set of genes selected for seed size also being implicated in shoot traits is not unexpected. Similarly, the ear number genes were identified from the Golden Glow selection experiment for ear number [36], where correlated changes in kernel size and kernel number were also observed [34].

Comparison of MeSH and GO overrepresentation analyses

Among the most obvious findings when comparing results from MeSH and GO for all five of the datasets is that the number of GO term associations dramatically surpasses the number identified by MeSH (Table 2). Within the sets of overrepresented terms (Additional files 1, 2, 3, 4, 5), there are cases of clearly overlapping GO and MeSH terms. For instance, in the improvement dataset, MeSH identified “Lipoxygenase” as the most significantly overrepresented term in the Chemicals and Drugs category, while GO identified the similar “linoleate 13S-lipoxygenase activity” term as highly significant in the Molecular Function category. However, there were instances where the MeSH analysis identified associations that were missed by GO. An example of this is that from the inflorescence dataset “Hybrid Vigor” was an enriched term in the Phenomena and Processes MeSH category, while no similar terms were identified by GO in any category. Although these examples are anecdotal, they are only a minor subset of the complete lists provided by this analysis and available for further scrutiny (Additional files 1, 2, 3, 4, 5). We mention the examples to demonstrate that MeSH and GO can either differ remarkably in their findings or, in some instances, particularly for highly significant terms, provide an independent confirmation that the other method is on the right track.

The most meaningful difference between MeSH and GO analyses is the source from which the annotations are derived. While most GO annotations are assigned algorithmically [37] with little or no human input [38], MeSH annotations are derived from manually curated manuscript classifications. This difference seems to lead to the existence of MeSH terms that correspond to easily interpretable macro-scale phenotypes, but it introduces additional complications as well. For example, the mention of a specific gene in a manuscript about hybrid-vigor may lead to a MeSH annotation of “hybrid-vigor” for that gene, even if no direct link was implied by the authors. However, this is a consideration that should always be at the forefront of ORA, regardless of the annotation scheme being used. To summarize, since MeSH and GO analyses are based on wholly different annotation mechanisms, the two approaches have the potential complement one another nicely. It is not our intention to suggest that MeSH should supplant GO, or even be viewed as a competitor to GO, since both platforms can provide distinct insights.

Current limitations

Despite the promising MeSH ORA and semantic similarity results observed in this study, using MeSH to guide biological interpretations still has an assortment of limitations that should be considered during any study that involves MeSH. Firstly, for non-model organisms,

including maize and other crops, relatively few genes have corresponding manuscripts that have been directly annotated with MeSH terms. Additionally, due to the nature of NCBI-based annotations, a requirement of current software is that all genes have Entrez Gene ID's [39] to enable mapping from genes to MeSH terms, but Entrez Gene ID's have only been assigned to a subset of maize genes. In fact, among the five datasets we analyzed, approximately two-thirds of the genes falling within the putatively functional regions did not have a corresponding Entrez Gene ID. This is particularly troubling in light of our observations regarding the ear number gene set, which was the smallest list of genes considered. Only 195 genes were contained within the selected regions (compared to thousands for some of the other data sets), and only 62 of those had corresponding Entrez Gene IDs. With fewer genes included during ORA, the power to detect significant enrichment is reduced. Similarly, this dataset showed very weak similarity to the others, which we hypothesize is at least in part due to the limited number of included genes and corresponding MeSH terms.

Conclusions

Even considering the above limitations, we expect MeSH-based analyses will improve over time. As additional mapping and functional manuscripts are published, the number of Entrez genes and the descriptive MeSH terms corresponding to each, in both model and non-model species, will increase. This increase will improve the magnitude and reliability of results gleaned from MeSH. Although improvements are expected with time, the five datasets studied here demonstrate how MeSH can currently be leveraged for making biological interpretations in maize as well as other crop and plant species.

Additional files

Additional file 1 R-Markdown file including script and results of MeSH and GO analysis on maize domestication genes.

Additional file 2 R-Markdown file including script and results of MeSH and GO analysis on maize improvement genes.

Additional file 3 R-Markdown file including script and results of MeSH and GO analysis on maize genes under selection for an increase in maize seed size.

Additional file 4 R-Markdown file including script and results of MeSH and GO analysis on maize genes under selection for an increase in ear number per plant.

Additional file 5 R-Markdown file including script and results of MeSH and GO analysis on maize genes implicated in a GWAS study of maize inflorescence traits.

Additional file 6 R-Markdown file including script and results of MeSH and GO analysis on a random set of 1500 maize genes.

Additional file 7 R-Markdown file including script and results of MeSH semantic similarity analysis.

Abbreviations

GWAS: genome wide association study; MeSH: Medical Subject Headings; NLM: National Library of Medicine; ORA: overrepresentation analysis.

Authors' contributions

TB and GM conceived the study. TB conducted the analysis with substantial input from GM. TB generated the additional R-markdown files. Both authors read and approved the final manuscript.

Author details

¹ USDA-ARS Plant Genetics Research Unit, Division of Plant Sciences, Division of Biological Sciences, MU Informatics Institute, University of Missouri, Columbia, MO 65211, USA. ² Department of Animal Science, University of Nebraska, Lincoln, NE 68583, USA.

Acknowledgements

We wish to thank the authors of the five previously-published manuscripts that released the publicly available datasets which made this study possible.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All scripts generated and used for study are available in additional annotated R-markdown files (Additional files 1, 2, 3, 4, 5). These files are designed to facilitate implementation by other researchers. All data analyzed are publicly available from the references provided in-text.

Funding

This work was supported by the USDA Agricultural Research Service. A University of Nebraska Layman fund provided support for Gota Morota.

Received: 2 October 2016 Accepted: 18 February 2017

Published online: 23 February 2017

References

- Ogura T, Busch W. From phenotypes to causal sequences: using genome wide association studies to dissect the sequence basis for variation of plant development. *Curr Opin Plant Biol*. 2015;23:98–108.
- Lorenz AJ, Beissinger TM, Rodrigues R, de Leon N. Selection for silage yield and composition did not affect genomic diversity within the Wisconsin Quality Synthetic maize population. *Genes Genomes Genet*. 2015; doi:10.1534/g3.114.015263.
- Rausher MD, Delph LF. Commentary: When does understanding phenotypic evolution require identification of the underlying genes? *Evolution*. 2015;69:1655–64.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. *Nat Genet*. 2000;25:25–9.
- Balakrishnan R, et al. A guide to best practices for gene ontology (GO) manual annotation. *Database* 2013;2013:bat054. doi:10.1093/database/bat054.
- Gene Ontology Consortium, et al. Gene ontology annotations and resources. *Nucleic Acids Res*. 2013;41:D530–5.
- Kanehisa M, Goto S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27–30.
- Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res*. 2012;40:D940–6.
- Lipscomb CE. Medical Subject Headings (MeSH). *Bull Med Libr Assoc*. 2000;88:265.
- PUBMED Medical Subject Headings. <https://www.nlm.nih.gov/mesh/meshhome.html>. Accessed Sept 2016.
- Nakazato T, Takinaka T, Mizuguchi H, Matsuda H, Bono H, Asogawa M. Bio-compass: a novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes. *In Silico Biol*. 2008;8:53–61.
- R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015. <https://www.R-project.org/>.
- Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oles AK, Pages H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015;12:115–21.
- Tsuyuzaki K, Morota G, Ishii M, Nakazato T, Miyazaki S, Nikaido I. Mesh ora framework: R/Bioconductor packages to support mesh over-representation analysis. *BMC Bioinform*. 2015;16:45.
- Morota G, Beissinger TM, Peñagaricano F. MeSH-informed enrichment analysis and MeSH-guided semantic similarity among functional terms and gene products in chicken. *Genes Genomes Genet*. 2016;6:2447–53.
- Morota G, Peñagaricano F, Petersen JL, Ciobanu DC, Tsuyuzaki K, Nikaido I. An application of MeSHF enrichment analysis in livestock. *Anim Genet*. 2015;46:381–7.
- Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppeler SM, et al. Comparative population genomics of maize domestication and improvement. *Nat Genet*. 2012;44:808–11.
- Hirsch CN, Flint-Garcia SA, Beissinger TM, Eichten SR, Deshpande S, Barry K, McMullen MD, Holland JB, Buckler ES, Springer N, et al. Insights into the effects of long-term artificial selection on seed size in maize. *Genetics*. 2014;198:409–21.
- Beissinger TM, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, Buell CR, Kaeppeler SM, Gianola D, de Leon N. A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. *Genetics*. 2014;196:829–40.
- Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S, Holland JB, Flint-Garcia S, McMullen MD, Buckler ES, et al. Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet*. 2011;7:e1002383.
- Zhou J, Shui Y. MeSHSim: MeSH (Medical Subject Headings) semantic similarity measures. 2015; R package version 1.2.0.
- gene2pubmed. <ftp://ftp.ncbi.nih.gov/gene/DATA>. Accessed July 2016.
- Falcon S, Gentleman R. Using GOSTats to test gene lists for GO term association. *Bioinformatics*. 2007;23:257–8.
- Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 2009;4:1184–91.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. The b73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326:1112–5.
- Ensembl Plants <http://plants.ensembl.org/index.html>. Accessed July, 2016.
- Wei T. corrplot: visualization of a correlation matrix. 2013; R package version 0.73.
- Pavlidis P, Jensen JD, Stephan W, Stamatakis A. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Mol Biol Evol*. 2012;29:3237–48.
- Gottlieb LD. Genetics and morphological evolution in plants. *Am Nat* 1984;123(5):681–709.
- Dorweiler J, Doebley J. Developmental analysis of teosinte glume architecture1: a key locus in the evolution of maize (poaceae). *Am J Bot*. 1997;84:1313.
- Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, Bombliès K, Lukens L, Doebley JF. The origin of the naked grains of maize. *Nature*. 2005;436:714–9.
- Wang H, Studer AJ, Zhao Q, Meeley R, Doebley JF. Evidence that the origin of naked kernels during maize domestication was caused by a single amino acid substitution in tga1. *Genetics*. 2015;200:965–74.
- Doebley J, Stec A, Hubbard L. The evolution of apical dominance in maize. *Nature*. 1997;386:485–8.
- De Leon N, Coors J. Twenty-four cycles of mass selection for prolificacy in the golden glow maize population. *Crop Sci*. 2002;42:325–33.
- Sekhon RS, Hirsch CN, Childs KL, Breitzman MW, Kell P, Duvick S, Spalding EP, Buell CR, de Leon N, Kaeppeler SM. Phenotypic and transcriptional analysis of divergently selected maize populations reveals the role of developmental timing in seed size determination. *Plant Physiol*. 2014;165:658–69.
- Maita R, Coors J. Twenty cycles of biparental mass selection for prolificacy in the open-pollinated maize population golden glow. *Crop Sci*. 1996;36:1527–32.

37. du Plessis L, Škunca N, Dessimoz C. The what, where, how and why of gene ontology—a primer for bioinformaticians. *Brief Bioinform.* 2011;12(6):723–35. doi:[10.1093/bib/bbr002](https://doi.org/10.1093/bib/bbr002).
38. Škunca N, Altenhoff A, Dessimoz C. Quality of computationally inferred gene ontology annotations. *PLoS Comput Biol.* 2012;8:1–11. doi:[10.1371/journal.pcbi.1002533](https://doi.org/10.1371/journal.pcbi.1002533).
39. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.* 2005;33:D54–8.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

