iSyTE: integrated Systems Tool for Eye gene discovery

Salil A. Lachke,[1,2,9] Joshua W.K. Ho,[1,3,9] Gregory V. Kryukov,[1,4,9] Daniel J. O'Connell,[1] Anton Aboukhalil,[1,5] Martha L. Bulyk,[1,6,7] Peter J. Park,[1,3,8] Richard L. Maas[1]*

Affiliations:

[1] Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115 USA

[2] Department of Biological Sciences and Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19716 USA

[3] Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115 USA

[4] The Broad Institute, Cambridge, MA 02139 USA

[5] Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

[6] Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115 USA

[7] Harvard-Massachusetts Institute of Technology Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA

[8] Informatics Programs, Children's Hospital of Boston, Boston, MA 02115, USA

[9] Authors contributed equally

* **Correspondence:** maas@genetics.med.harvard.edu (R.L.M.)

Word Count of Text: 3722

Grant Information: This work was supported by NIH grants R01EY10123-15 to RLM and R01EY021505-01 to SAL. JWKH was supported by the NIH Common Fund via NIBIB RL9EB008539.

ABSTRACT

PURPOSE: To facilitate the identification of genes associated with cataract and other ocular defects, we developed and validated a computational tool termed *iSyTE* (integrated Systems Tool for Eye gene discovery; http://bioinformatics.udel.edu/Research/iSyTE). *iSyTE* uses a mouse embryonic lens gene expression dataset as a bioinformatic filter to select candidate genes from human or mouse genomic regions implicated in disease, and to prioritize them for further mutational and functional analyses.

METHODS: We obtained microarray gene expression profiles for microdissected embryonic mouse lens at three key developmental timepoints as it transitions from the embryonic day E10.5 stage of lens placode invagination to E12.5 lens primary fiber cell differentiation. Differentially regulated genes were identified by *in silico* comparison of lens gene expression profiles to those of whole embryo body (WB) lacking ocular tissue.

RESULTS: Gene set analysis demonstrates that this strategy effectively removes highly expressed but non-specific housekeeping genes from lens tissue expression profiles, allowing identification of less-highly expressed lens disease associated genes. Among 24 previously mapped human genomic intervals containing genes associated with isolated congenital cataract, the mutant gene is ranked within the top 2 *iSyTE* selected candidates in ~88% cases. Finally, *in situ* hybridization confirmed lens expression of several novel *iSyTE*-identified genes.

CONCLUSIONS: *iSyTE* is a publically available web resource that can be used to prioritize candidate genes within mapped genomic intervals associated with congenital cataract for further investigation. Extension of this approach to other ocular tissue components will facilitate eye disease gene discovery.

KEYWORDS

Cataract, Lens, Microarray, Structural Birth Defect, Systems Biology, Tooth, Organogenesis,

Even with the advent of high-throughput sequencing, the discovery of genes associated with congenital birth defects such as eye defects remains a challenge. We sought to develop a straightforward experimental approach that could facilitate the identification of candidate genes for developmental disorders, and as proof-of-principle, we chose defects involving the ocular lens. Opacification of the lens results in cataract, a leading cause of blindness that affects 77 million individuals and accounts for 48% of the blind.[1] Cataracts can be classified as either congenital or age-related, and can be expressed as either an "isolated", "non-syndromic" phenotype, or as part of a larger developmental syndrome.[2-4] Approximately one quarter of congenital cataracts are inherited,[5] and all three modes of Mendelian inheritance have been described, with autosomal dominant being the most common.[2] Both linkage and mutational analyses of candidate genes have been successfully used to identify genetic causes of congenital cataracts, and presently 24 loci exists for isolated cataracts.[2]

The identification of genetic mutations - such as those implicated in cataract formation - traditionally follows an initial mapping step that involves linkage analysis or homozygosity mapping, followed by sequence analysis of candidate genes or genomic regions in patient DNA. A similar approach can identify mutant genes in model organisms such as mouse and zebrafish. Nonetheless, linkage and mutational analysis are cumbersome, and often involve the exclusion of a large number of candidate genes by DNA sequence analysis before the correct gene is identified. While the advent of next generation sequencing makes it possible to rapidly identify a large number of potentially deleterious genetic variants within a sample, it often remains unclear how to identify the actual disease-associated mutation in a cost-effective manner without performing a large cohort case-control study. It is often the case that additional biological knowledge is necessary to resolve disease-producing genetic mutations from sequence variants that are unrelated to the phenotype of interest.

In the case of human developmental disorders, we hypothesized that knowledge of embryonic gene expression patterns, which are often conserved and readily accessible for the homologous mouse genes, could help assist in the identification of congenital birth defect genes in human. Here we describe a straightforward experimental and computational strategy to identify and prioritize candidate disease genes based on microarray gene expression profiles that are generated from embryonic mouse tissues. As an initial application, we applied this to cataract phenotypes. To make this tool broadly accessible, we concurrently developed a publicly available web-based resource termed *iSyTE* (integrated Systems Tool for Eye gene discovery) that can efficiently prioritize candidate genes associated with human congenital cataract.

METHODS

Mouse Husbandry

Mice were treated in accordance with protocols established by the Association for Research in Vision and Ophthalmology (ARVO). The Animal Care and Use Committee (IACUC) of Harvard Medical School (Boston, MA) approved all experimental protocols involving mice. Wild type ICR mice were obtained from Taconic (Albany, NY) and used for microarray and *in situ* hybridization analyses. Mice were housed in a 14 h light and 10 h dark cycle, and the morning of vaginal plug discovery was defined as embryonic day E0.5.

Microarray Analysis

Total RNA was extracted from manually dissected mouse embryonic day 10.5, 11.5 and 12.5 lenses (approximately 200 lenses per E10.5 replicate, 150 lenses per E11.5 replicate, 100 lenses per E12.5 replicate), or from whole embryonic tissue minus the eye region at stages E10.5, E11.5 and E12.5 using the RNeasy Mini Kit (Qiagen). RNA from stage matched whole embryonic

tissue minus the eye region, which was removed by microdissection, was pooled in equimolar ratios, denoted the whole body or "WB" control, and was processed in parallel. Microarray data from the WB control was later used to achieve *in silico* enrichment for lens enriched genes (see Results). We first tested the purity of the dissected lens tissue by analyzing dissected lenses at these stages from *P0-3.9-GFPCre* reporter mice, in which the lens-specific GFP expression is driven by the *Pax6* ectodermal enhancer within the 3.9-kb region upstream of the *Pax6* P0 promoter.[6] We then used wild-type in house timed pregnant ICR mice as a resource for collecting the lens tissues that were used for microarray analysis. Microarray analyses were performed in biological triplicate by hybridization to the Affymetrix Mouse 430 2.0 chip in the Biopolymers facility at Harvard Medical School. Standard Affymetrix protocols were used to prepare cDNA and biotin labeled cRNA using *in vitro* transcription. Quality of the total RNA was evaluated via an Agilent 2100 Bioanalyzer prior to being processed for cDNA preparation by RT-PCR. The cDNA was converted to biotinylated cRNA using modified NTPs in an *in vitro* transcription reaction. The labelled cRNA was hybridized to the chips for 16 hr and then washed and stained. The chip was irradiated at 488 nm (excitation) and scanned at 570 nm (emission). Raw probe intensities from all microarray profiles were preprocessed together using the robust multiarray average (RMA) method,[7] implemented in the *affy* package.[8] If a gene was represented by multiple probe sets, we selected the probe set with the highest median expression across all samples to represent the expression of that gene. In this manner, all probe sets were collapsed into 20,460 genes, based on their unique gene symbols. To calculate tissue specific enrichment, we used a moderated *t*-test implemented in *limma*[9] to identify differentially expressed genes. False discovery rates (FDRs) were then estimated for this gene list using the method of Benjamini and Hochberg.[10] All bioinformatics analyses were carried out using an R

statistical environment (http://www.r-project.org). The NCBI Gene Expression Omnibus accession number for all the microarray data reported in this paper is GSE32334.

Gene set analysis

To perform a comprehensive and unbiased gene set analysis, we utilized a large compendium of over 10,000 mouse specific gene sets comprising Gene Ontology terms, KEGG pathways, MouseCyc pathways, MGI mouse phenotype associated genes, FANTOM4 mouse tissue specific transcription factor gene sets, and other custom gene sets related to development, signaling pathways, and stem cell regulation (Supplementary Table S1). Furthermore, we compiled gene sets for lens development, human cataract, and for control purposes, tooth development, human tooth agenesis, and human orofacial clefting (Supplementary Table S2). For lens development genes, we used a recently curated a list of genes that are critically involved in the pre-placodal and placodal stages of lens development.[11] In addition, we compiled lists of non-syndromic and syndromic human cataract genes based on a high quality manual collection of all known human cataract associated genes, Cat-Map.[2] Tooth development genes, for comparative purposes, were those that cause abnormal tooth development in mouse and/or human models, based on the Mouse Genome Informatics (MGI) database (mammalian phenotype ID: MP:0000116). Similarly, tooth agenesis and orofacial clefting gene lists were taken from a recent review,[12] with the addition of one new non-syndromic tooth agenesis gene: Wnt10a.[13] Full details of these gene sets are available in Supplementary Table S2. We tested whether the 200 most highly ranked genes (with or without WB control) were enriched for each gene set independently using Fisher's exact test. The resulting *p*-values were Bonferroni corrected.

***In situ* hybridization**

In situ hybridization experiments were performed as previously described.[14] In brief, primers containing SP6 or T7 promoter sequences upstream of gene-specific sequences were used to amplify cDNA products that were then analyzed by 1% agarose gel electrophoresis, column purified and used as templates in *in vitro* transcription UTP-digoxigenin labeling reactions. Digoxigenin-labeled probes were then used for *in situ* hybridization on 13 μm E11.5 mouse embryonic lens frozen sections. Primer pairs 5′-GCTATTTAGGTGACACTATAGTCT-ACCTGGGCTTTCTGGTG-3′, *Fam198b*-F and 5′-TTGTAATACGACTCACTATAGGGGCA-TTCTGCGGATGTCTTCT-3′, *Fam198b*-R; Primers 5′-GCTATTTAGGTGACACTATAGT-CTCAGCTCCCAGCTTTGAT-3′, *Ptpru*-F and 5′-TTGTAATACGACTCACTATAGGGCTTT-GCGGATGATGACAATG-3′, *Ptpru*-R; Primers 5′-GCTATTTAGGTGACACTATAGA-GCTTCACCCAGCCCTTATC-3′, *Ng23*-F and 5′-TTGTAATACGACTCACTATAGGGTCT-GTCTGCAGCTGTTGAGG-3′, *Ng23*-R; Primers 5′-GCTATTTAGGTGACACTATAGGA-CCATCGAGGACGACCTAA-3′, *Sipa1l3*-F and 5′-TTGTAATACGACTCACTATAGGG-GAGTGGCTCTTGGAGTCTGG-3′, *Sipa1l3*-R; Primers 5′-GCTATTTAGGTGACACTATAG-TACCTACCCTCCTGCCACAG-3′, *Ypel2*-F and 5′-TTGTAATACGACTCACTATAGGG-CCCAAAGTGGTTTTGCAGTT-3′, *Ypel2*-R; Primers 5′-GCTATTTAGGTGACACTATA-GGAATCATGCAGCCAGGTTTT-3′, *Rbm24*-F and 5′-TTGTAATACGACTCACTATAGGG-TCTGTCTGCAGCTGTTGAGG-3′, *Rbm24*-R; Primers 5′-GCTATTTAGGTGACACTATA-GGGCCAGTTCCACACTCTCTT-3′, *Gje1*-F and 5′-TTGTAATACGACTCACTATAGGG-CTCAAAAACCTCAGCAACACA-3′, *Gje1*-R; Primers 5′-GCTATTTAGGTGACACTATA-GGACACAGGCTCAAGCTACCC-3′, *Vit*-F and 5′-TTGTAATACGACTCACTATAGGG-CCATTGGCTTTGGAAAAGAA-3′, *Vit*-R; were used to amplify mRNA-specific probe

sequence from E12.5 mouse embryonic cDNA. Digitized images were processed using Adobe Photoshop. Reagents and probes are available upon request.

RESULTS

Gene expression profiling of the mouse embryonic lens

To construct the *iSyTE* database, we identified three critical time points in lens development, at E10.5, E11.5 and E12.5, as the lens transitions from the stage of lens placode invagination (E10.5) to that of lens vesicle formation and the onset of lens fiber cell differentiation (E12.5) (Fig. 1).[11,15] This developmental window conforms to when mouse orthologs of many human cataract genes are strongly expressed in the developing mouse lens. To ensure high quality microarray data, we isolated total RNA from manually micro-dissected mouse embryonic lenses at these stages in amounts sufficient to use a single step cDNA amplification protocol (see Methods). Using whole genome transcript profiling on Affymetrix Mouse Genome 430 2.0 microarrays, we generated a developmental profile of the mouse lens transcriptome over the specified developmental interval. The quality of the processed microarrays was assessed using various diagnostic plots, and no anomalies were found (Supplementary Fig. S1).

Identification of lens-enriched genes

To identify genes with lens-enriched expression, we established an *in silico* subtraction approach by which lens microarray datasets are compared to a developmentally matched microarray dataset representing the whole embryonic body from which the ocular tissue was removed by microdissection, denoted "WB". This *in silico* subtraction involves ranking all genes based on *t*-statistic when tissue specific expression profiles are compared to WB profiles. We hypothesized that this control background dataset, which we denoted "WB" for "whole body

minus eyes", represents an optimal averaged gene expression profile for a mixture of tissues, and that comparison of tissue-specific profiles against the WB control profile would facilitate identification of genes with lens-specific or lens-enriched expression. We anticipated that the resulting *in silico* subtracted mouse lens database would represent a useful tool to identify lens-enriched genes with roles in lens biology, and with which to prioritize candidate genes within mapped cataract loci for mutational analysis. While exceptions exist, this is consistent with the hypothesis that tissue enriched gene expression more likely reflects a function for the gene in that tissue, than if a gene exhibits ubiquitous or widespread expression. The ranked lists of lens enriched genes are what we refer to as the *iSyTE* database.

We tested the utility of this approach to identify genes associated with lens development and human cataract by first identifying the gene sets that are enriched in the top 200 highly ranked genes (representing ~1% of total number of genes in the genome) - with or without WB control - using Fisher's exact test with Bonferroni corrected *p*-values. The top 200 highly ranked genes from the lens dataset with WB subtraction were highly enriched for gene sets for eye and lens biology, without being enriched for gene sets for miscellaneous housekeeping factors (Fig. 2A). We also identified the most highly enriched gene sets for the top 200 highly ranked genes from the lens dataset without WB subtraction, and found that they mainly consisted of ribosomal proteins. Therefore, the *in silico* subtraction method specifically identifies lens-enriched genes - both with high expression and low expression in the lens - while filtering out genes with high expression that are not lens-specific. We further found that the top 200 lens-enriched genes from the WB subtraction dataset consist mainly of genes associated with lens development, isolated or non-syndromic cataract, and interestingly, with syndromic cataract as well (Fig. 2B, C; Supplementary Fig. S2). Analysis using different numbers of top lens-enriched genes (such as *n*=100, 300, 500 genes) produced similar results (data not shown).

***iSyTE* effectively identifies known and novel genes associated with cataract**

To test the potential of *iSyTE* to identify cataract associated genes, we analyzed 24 previously mapped intervals that contain genes associated with human isolated or non-syndromic congenital cataract. Upon manual inspection of these mapped genomic intervals, *iSyTE* successfully identified the correct mutant gene as the top candidate within a locus in ~70% cases (17/24), and in ~88% cases (21/24) it ranked the mutant gene within the top 2 candidates among all candidate genes in the locus, where each locus spans on average 12.3 Mb and contains about 80 genes (Table 1). Moreover, the effectiveness of mutant gene identification remained high even when the highly lens-specific crystallin encoding genes were removed from the analysis. These data reflect the ability of *iSyTE* to identify genes that are expressed at relatively low levels, but that are highly enriched in the lens. This group includes the genes *FOXE3*, *HSF4*, *MAF* and *PITX3* that encode transcription factors, as well as *BSFP2*, *LIM2* and *MIP* that encode cytoskeletal proteins (Table 2).

In addition to the identification of known cataract genes, *iSyTE* can also identify novel cataract genes. We successfully employed a preliminary version of *iSyTE* to identify the genes involved in two separate cataract cases.[16,17] In the first case, the patient presented with bilateral, progressive cataracts with posterior lenticonus as the primary phenotype and carried the balanced paracentric inversion 46,XY,inv(9)(q22.33q34.11).[16] The *iSyTE* database identified *TDRD7* as the most probable candidate among 110 genes within a 10 Mb interval around the q22.33 breakpoint. Subsequently, disruption and haploinsufficiency of *TDRD7* in the patient was confirmed, and an additional independent 3 bp coding region deletion mutation in *TDRD7* was identified in a consanguineous case. In the second case, we applied *iSyTE* to another independent case of human congenital cataract, in which a translocation breakpoint ostensibly

responsible for the proband's phenotype was located within a relatively gene-poor genomic interval in which no gene was directly interrupted.[17] Nonetheless, *iSyTE* correctly identified *PVRL3* as the gene responsible for the proband's cataract phenotype, most likely on the basis of a position effect, as subsequently proven by the analysis of multiple mouse *Pvrl3* mutant alleles.

As yet another validation of *iSyTE*, we used section *in situ* analysis for several *iSyTE*-identified genes on mouse embryonic lens sections to confirm that some of the novel genes that *iSyTE* ranked as lens-enriched were indeed expressed in the expected fashion (Fig. 3). This analysis demonstrated highly enriched lens expression of all eight of eight randomly chosen genes that were ranked within the top 250 lens-enriched genes, establishing the validity of the database (Fig. 3). Moreover, human orthologs of two of these genes (*SIPA1L3* and *PTPRU*) fall within or near mapped human cataract loci.[18,19] *Gje1* (previously known as: *Gjf1*; Fig. 3) has been recently identified as a novel cataract-associated gene in a mouse model.[20] Besides these eight relatively uncharacterized genes, evidence for lens enrichment and association with cataract in mouse models has also recently been documented for other *iSyTE* lens-enriched genes, *e.g. Aldh1a1*.[21] These results further support the utility of *iSyTE* as a cataract gene prioritization resource.

We next sought to use *iSyTE* to predict promising candidate genes in mapped human cataract loci for which the gene involved has not been identified. We analyzed the latest version of the Cat-Map dataset[2] (latest update September 30, 2011) and identified 17 mapped cataract intervals for which a gene has not yet been assigned. We then used *iSyTE* to predict the most promising candidate genes in these loci. We provide the top candidate genes in each mapped interval based on their high lens-enrichment rank in *iSyTE* (Table 3; Supplementary Table 3). Based on our result that 88% (21/24) of known cataract genes are within the top 2 candidate genes within a mapped interval, the gene list in Table 3 of *iSYTE*-predicted candidate cataract

genes can potentially serve as a resource for identifying and prioritizing cataract associated candidate genes for sequencing.

Basis of the effectiveness of the subtraction strategy

To understand the basis for the effectiveness of the subtraction strategy in identifying genes of functional significance in lens development, we compared gene expression between the developing lens and WB control (Table 2). As expected, dramatic differences for signal intensities of genes encoding crystallin proteins between the lens and WB control were observed. However, genes with relatively low levels of expression in the lens microarray database, which otherwise would likely be ranked as low priority candidates (*e.g., Hsf4*, *Bfsp2*), are identified by the subtraction strategy; genes encoding developmental transcription factors also appear to be preferentially selected.

Lastly, the microarray expression patterns in the three developmental stages appears to faithfully reflect the published expression pattern of genes in lens development. For example, *Bmp7*, *Meis1*, *Sox2*, *Pax6* and *Mab21l1*, which function in early lens development, have progressively decreased expression by microarray from E10.5 through E12.5. In contrast, *Gja3*, *Gja8*, *Sox1*, *Prox1*, *Mip* and *Lim2*, which function in lens fiber cells, have progressively increased expression by microarray from E10.5 through E12.5. Thus, because of its derivation from three temporally distinct stages of lens development, the *iSyTE* database provides insight into early or late function for the gene of interest.

Extension of the subtraction strategy to other tissue types

To investigate whether the *in silico* subtraction strategy could be generally applied to identify genes associated with other developmental disorders, we generated a microarray dataset for the

developing molar tooth, which is a well-established system for studying the epithelial-mesenchymal interactions involved in organogenesis. We performed laser capture microdissection (LCM) to capture mouse embryonic E13.5 tooth germ tissue and then extracted sufficient total RNA to perform microarrays after two rounds of *in vitro* transcription-based amplification (double amplification) (Supplementary Fig. 3). Using the same amplification protocol, we also generated a microarray dataset from total RNA extracted and pooled in equimolar ratios from mouse whole body (WB) tissue at E11.5, E12.5 and E13.5. Similar to the lens, the tooth specific profiles were "subtracted" from the WB control using a moderated *t*-test and a tooth enrichment *p*-value was assigned to each gene. *t*-statistics were used to rank genes for tooth enrichment.

We next tested the utility of this strategy to identify genes associated with tooth development and human tooth and craniofacial defects. Similar to the lens, these analyses demonstrate that the top 200 highly ranked genes after WB subtraction were highly enriched for genes relevant to tooth biology, without being enriched for genes encoding miscellaneous house keeping factors (Supplementary Fig. 4A). As expected, the top 200 tooth-enriched genes from the WB subtraction dataset contained genes associated with syndromic and non-syndromic tooth agenesis and with orofacial clefting (Supplementary Fig. 4B, C). These data are accessible at: http://bioinformatics.udel.edu/Research/iSyTE and indicate that in addition to the lens, the *in silico* subtraction strategy can successfully identify genes associated with tooth development and disease.

Use of WB microarray datasets as a public resource

We next sought to test the robustness and applicability of the two different WB datasets generated in this study. The two sets of WB microarray gene expression profiles generated in

this study were generated by two experimentalists, using two different amplification protocols (single amplification for lens, and double amplification for tooth), and at slightly different developmental stages (E10.5, 11.5 and 12.5 for lens, and E11.5, 12.5 and 13.5 for tooth). We tested whether we could still identify tissue specific gene enrichment even when WB profiles were generated from a different preparation. Indeed, swapping the different WB profiles generated for the lens and the tooth analysis in the *in silico* subtraction strategy still robustly identified genes associated with lens and tooth developmental disorders, respectively (Fig. 4).

Construction of a web-based public resource*: iSyTE*

Finally, we sought to represent the lens enrichment data in user-friendly Genome Browser tracks, allowing our genome-wide lens enrichment data to be visualized in the context of the vast amount of genomic annotation that is already available. We created a custom *iSyTE* track at the University of California at Santa Cruz (UCSC) Genome Browser, and it is accessible from http://bioinformatics.udel.edu/Research/iSyTE. Each track is color-coded to represent the lens-enrichment ranking based on WB-subtracted gene expression profiles of E10.5, E11.5 and E12.5 lens. Thus, *iSyTE* tracks allow the visualization of genes with their degree of enrichment of expression in the developing lens expressed in a color-coded format, with red being highly enriched and blue highly depleted (Fig. 5). To make our resource useful for a wide variety of users, we provide *iSyTE* custom tracks for two widely used human genome assemblies (hg19 and hg18), and two mouse genome assemblies (mm9 and mm8).

Operationally, after opening the UCSC Genome Browser for a specific genome assembly the user can search for and browse any genomic interval of interest. This representation allows immediate visual detection of the best candidate genes in a given genomic interval, and allows one to zoom in or out to visualize the presence of promising candidates within a particular region

or proximal to it. The *iSyTE* tracks can be viewed in the context of other genomic resources that are already available in the UCSC Genome Browser, such as sequence conservation, known SNP locations, and ENCODE histone modification profiles. Lastly, visualization of the *iSyTE* tracks that represent three embryonic stages in one frame provides some appreciation of the dynamic pattern of gene expression during lens development.

DISCUSSION

Although it has been proposed that tissue-specific gene expression profiling may facilitate disease gene identification,[22,23] and gene expression datasets for many tissue and cell types exist, the application of these resources to gene discovery, particularly in the context of disease, has been limited.[24] This is largely because such datasets are large and the route to efficient selection and prioritization of candidate genes is not straightforward, especially in the context of normal development and in the absence of clear control versus mutant gene expression change comparisons. Several gene expression atlases that are based on *in situ* hybridization provide insight into developmental gene expression,[25] but such information is typically non-quantitative and does not permit facile comparison of tissue specific gene expression levels. In this work, we developed a strategy to subject tissue-specific microarray datasets to an *in silico* subtraction that involve a comparison of a tissue specific dataset with a whole body ("WB") reference dataset, which allows the systematic ranking of genes based on their tissue enrichment. Even with high throughput sequencing, mutations that lie outside of coding regions may be difficult to identify. We demonstrate that this filter provides a highly effective way to identify candidate genes associated with the development of specific tissues for which gene expression profiles can be readily obtained.

The development of *iSyTE* was based on two basic hypotheses: (1) genes that are highly expressed at critical stages of murine embryonic development in a specific organ are likely associated with mutations in human genes that are linked to an organ specific birth defect; and (2) *in silico* subtraction of gene expression profiles for whole embryonic body from those for equivalently staged specific, microdissected embryonic tissue can effectively remove non-specific but highly expressed genes, thereby revealing tissue specific genes. Using lens and tooth as examples, we show that this relatively straightforward experimental and computational approach can effectively facilitate the identification of human disease-associated genes.

As with any gene prediction tool, there is a false negative rate associated with a given prediction, and it is important to consider the potential source of false negatives when interpreting results from *iSyTE*. Our retrospective analysis of 24 known cataract genes indicates that ~10% of the genes do not have high lens expression or enriched expression as measured in the current microarray data, hence suggesting a false negative rate of about 10%. This could potentially result from: (1) the sensitivity of the microarray probes for these genes may be poor; (2) the expression of these genes may be restricted to a different developmental stage(s) than those analyzed; and (3) the effect of lens-specific expression is masked by neighboring genes within the candidate interval that have higher levels of lens-specific expression but which are non-causative.

Indeed, such examples are evident in our present data analysis. For example, in 3 cases out of 24 (*FYCO1*, *GCNT2*, *CHMP4B*), *iSyTE* did not rank the correct gene within the top 2 candidates in the interval (Table 1). On further analysis, in case of *FYCO1* (ranked 21/191), the mapped interval was large (12.21 Mb) containing 191 candidate genes, several of which exhibited significantly higher lens enriched expression than *FYCO1*. In case of *GCNT2* (ranked 7/21 within a 5.26 Mb interval), we find very low expression of this gene in the microarrays,

indicative of either sub-optimal probe binding or genuinely low expression at the lens stages analyzed. Lastly, in case of *CHMP4B* (ranked 34/43 in a 3.03 Mb mapped interval), this gene is significantly expressed in the lens (signal detection $p < 0.002$), but it is also significantly expressed in the WB control; as a result it does not have a high lens-enrichment rank, and is therefore not correctly identified by *iSyTE* as a likely candidate gene.

In some cases, *iSyTE* does not predict any promising candidate genes based on lens-enrichment (*e.g.,* in the mapped human cataract intervals on 2q33 and 17p24 (Table 3)). In yet another case (20p11.23-p12.1), *iSyTE* predicted *BFSP1* from 29 candidates in the interval (Table 3). However, in this interval, *BFSP1* has been sequenced and found to harboring no exonic or exon junction mutation, suggesting that the mutation resides in a regulatory region or in another gene. Therefore, in all cases, further experimental validation via mutational sequence analysis will be necessary in addition to the *in silico* predictions made by *iSyTE*.

Other genome-wide *in silico* analyses have recently been applied to the interpretation of candidate SNPs in genome-wide association studies (GWAS).[26] For example, Ernst and coworkers[27] showed that cell-type specific histone modification patterns can identify regulatory regions, and that knowledge of the location of these regulatory regions and their associated genes can aid in the interpretation of GWAS by providing potential regulatory mechanisms for each candidate SNP. Similarly, Ozkul and coworkers[28] have devised a strategy based on ChIP-seq data for the transcription factor CRX to rank candidate genes within mapped intervals for retinitis pigmentosa (RP). Combined with exome sequencing, this approach successfully identified a novel mutation in the gene *MAK*, which is associated with RP. In the work reported here, we demonstrate a cost-effective strategy to effectively prioritize mutations for human disease gene identification. Because embryonic dissections can be readily performed in many research laboratories, and because microarray is increasingly affordable, the *iSyTE* approach

should be applicable to other organ- and tissue specific diseases, as demonstrated by our tooth germ analysis.

In conclusion, we describe a novel strategy for identifying disease-associated genes that is supported by a publically available web resource called *iSyTE*. We recently used a preliminary version of *iSyTE* to help identify two human genes associated with cataract, *TDRD7* and *PVRL3*. Since there are likely many other candidate cataract associated genes that have not yet been identified, this web-based resource should provide a useful tool for the ocular genetics community. Besides serving to identify lens-specific disease genes, future versions of *iSyTE* that include expression datasets for other ocular components should further help identify additional genes that influence the development and biology of the eye.

SUPPLEMENTARY MATERIAL

Supplementary material includes four figures and three tables and can be found with this article online.

ACKNOWLEDGEMENTS

The authors thank Dr. Sung Choe for preliminary analysis and Dr. Shamil Sunyaev for helpful comments and Dr. Hongzhan Huang for help with hosting the website. The authors declare no conflicts of interest.

REFERENCES

1. Resnikoff S, Pascolini D, Etya'ale D, et al. Global data on visual impairment in the year 2002. *Bull World Health Organ.* 2004;82:844-851.
2. Shiels A, Bennett TM, Hejtmancik JF. Cat-Map: putting cataract on the map. *Mol Vis.* 2010;16:2007-2015.
3. Hejtmancik JF. Congenital cataracts and their molecular genetics. *Semin Cell Dev Biol.* 2008;19:134-14910.1016/j.semcdb.2007.10.003.
4. Graw J. Mouse models of cataract. *J Genet.* 2009;88:469-486.
5. Bermejo E, Martínez-Frías ML. Congenital eye malformations: clinical-epidemiological analysis of 1,124,654 consecutive births in Spain. *Am J Med Genet.* 1998;75:497-504.
6. Rowan S, Sigger T, Lachke SA, et al. Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. *Genes Dev.* 2010;24:980-985.
7. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003;4:249-264.
8. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics.* 2004;20:307-315.
9. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 2004;3:Article310.2202/1544-6115.1027.
10. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Statist Soc B (Methodological).* 1995;57:289-300.
11. Lachke SA, Maas RL. Building the developmental oculome: systems biology in vertebrate eye development and disease. *Wiley Interdiscip Rev Syst Biol Med.* 2010;2:305-323.
12. Matalova E, Fleischmannova J, Sharpe PT, Tucker AS. Tooth agenesis: from molecular genetics to molecular dentistry. *J Dent Res.* 2008;87:617-623.
13. Bohring A, Stamm T, Spaich C, et al. WNT10A mutations are a frequent cause of a broad spectrum of ectodermal dysplasias with sex-biased manifestation pattern in heterozygotes. *Am J Hum Genet.* 2009;85:97-105.
14. Xu PX, Woo I, Her H, Beier DR, Maas RL. Mouse Eya homologues of the Drosophila eyes absent gene require Pax6 for expression in lens and nasal placode. *Development.* 1997;124:219-231.
15. Donner AL, Lachke SA, Maas RL. Lens induction in vertebrates: variations on a conserved theme of signaling events. *Semin Cell Dev Biol.* 2006;17:676-685.
16. Lachke SA, Alkuraya FS, Kneeland SC, et al. Mutations in the RNA granule component TDRD7 cause cataract and glaucoma. *Science.* 2011;331:1571-1576.
17. Lachke SA, Higgins AW, Inagaki M, et al. The cell adhesion gene PVRL3 is associated with congenital ocular defects. *Hum Genet.* 2011;10.1007/s00439-011-1064-zAvailable: http://www.ncbi.nlm.nih.gov/pubmed/21769484.
18. Bateman JB, Richter L, Flodman P, et al. A new locus for autosomal dominant cataract on chromosome 19: linkage analyses and screening of candidate genes. *Invest Ophthalmol Vis Sci.* 2006;47:3441-3449.
19. Hattersley K, Laurie KJ, Liebelt JE, et al. A novel syndrome of paediatric cataract, dysmorphism, ectodermal features, and developmental delay in Australian Aboriginal family maps to 1p35.3-p36.32. *BMC Med Genet.* 2010;11:16510.1186/1471-2350-11-165.

20. Puk O, Löster J, Dalke C, et al. Mutation in a novel connexin-like gene (Gjf1) in the mouse affects early lens development and causes a variable small-eye phenotype. *Invest Ophthalmol Vis Sci.* 2008;49:1525-1532.
21. Lassen N, Bateman JB, Estey T, et al. Multiple and additive functions of ALDH3A1 and ALDH1A1: cataract phenotype and ocular oxidative damage in Aldh3a1(-/-)/Aldh1a1(-/-) knock-out mice. *J Biol Chem.* 2007;282:25668-25676.
22. Blackshaw S, Fraioli RE, Furukawa T, Cepko CL. Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell.* 2001;107:579-589.
23. Diehn JJ, Diehn M, Marmor MF, Brown PO. Differential gene expression in anatomical compartments of the human eye. *Genome Biol.* 2005;6:R7410.1186/gb-2005-6-9-r74.
24. Brown JD, Dutta S, Bharti K, et al. Expression profiling during ocular development identifies 2 Nlz genes with a critical role in optic fissure closure. *Proc Natl Acad Sci USA.* 2009; 106:1462-1467.
25. de Boer BA, Ruijter JM, Voorbraak FPJM, Moorman AFM. More than a decade of developmental gene expression atlases: where are we now? *Nucleic Acids Res.* 2009; 37:7349-7359.
26. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. *Bioinformatics.* 2011;27:1741-1748.
27. Ernst S, Kirchner S, Krellner C, et al. Emerging local Kondo screening and spatial coherence in the heavy-fermion metal YbRh2Si2. *Nature.* 2011;474:362-366.
28. Ozgül RK, Siemiatkowska AM, Yucel D, et al. Exome sequencing and cis-regulatory mapping identify mutations in MAK, a gene encoding a regulator of ciliary length, as a cause of retinitis pigmentosa. *Am J Hum Genet.* 2011;89:253-264.
29. Ramachandran RD, Perumalsamy V, Hejtmancik JF. Autosomal recessive juvenile onset cataract associated with mutation in BFSP1. *Hum Genet.* 2007;121:475-482.
30. Jakobs PM, Hess JF, FitzGerald PG, et al. Autosomal-dominant congenital cataract associated with a deletion mutation in the human beaded filament protein gene BFSP2. *Am J Hum Genet.* 2000;66:1432-1436.
31. Shiels A, Bennett TM, Knopf HLS, et al. CHMP4B, a novel gene for autosomal dominant cataracts linked to chromosome 20q. *Am J Hum Genet.* 2007;81:596-606.
32. Litt M, Kramer P, LaMorticella DM, et al. Autosomal dominant congenital cataract associated with a missense mutation in the human alpha crystallin gene CRYAA. *Hum Mol Genet.* 1998;7:471-474.
33. Berry V, Francis P, Reddy MA, et al. Alpha-B crystallin gene (CRYAB) mutation causes dominant congenital posterior polar cataract in humans. *Am J Hum Genet.* 2001;69:1141-1145.
34. Padma T, Ayyagari R, Murty JS, et al. Autosomal dominant zonular cataract with sutural opacities localized to chromosome 17q11-12. *Am J Hum Genet.* 1995;57:840-845.
35. Billingsley G, Santhiya ST, Paterson AD, et al. CRYBA4, a novel human cataract gene, is also involved in microphthalmia. *Am J Hum Genet.* 2006;79:702-709.
36. Mackay DS, Boskovska OB, Knopf HLS, Lampi KJ, Shiels A. A nonsense mutation in CRYBB1 associated with autosomal dominant cataract linked to human chromosome 22q. *Am J Hum Genet.* 2002;71:1216-1221.
37. Kramer P, Yount J, Mitchell T, et al. A second gene for cerulean cataracts maps to the beta crystallin region on chromosome 22. *Genomics.* 1996;35:539-542.

38. Riazuddin SA, Yasmeen A, Yao W, et al. Mutations in betaB3-crystallin associated with autosomal recessive cataract in two Pakistani families. *Invest Ophthalmol Vis Sci.* 2005;46:2100-2106.
39. Héon E, Liu S, Billingsley G, et al. Gene localization for aculeiform cataract, on chromosome 2q33-35. *Am J Hum Genet.* 1998;63:921-926.
40. Sun H, Ma Z, Li Y, et al. Gamma-S crystallin gene (CRYGS) mutation causes dominant progressive cortical cataract in humans. *J Med Genet.* 2005;42:706-710.
41. Zhang T, Hua R, Xiao W, et al. Mutations of the EPHA2 receptor tyrosine kinase gene cause autosomal dominant congenital cataract. *Hum Mutat.* 2009;30:E603-611.
42. Chen J, Ma Z, Jiao X, et al. Mutations in FYCO1 cause autosomal-recessive congenital cataracts. *Am J Hum Genet.* 2011;88:827-838.
43. Pras E, Raz J, Yahalom V, et al. A nonsense mutation in the glucosaminyl (N-acetyl) transferase 2 gene (GCNT2): association with autosomal recessive congenital cataracts. *Invest Ophthalmol Vis Sci.* 2004;45:1940-1945.
44. Mackay D, Ionides A, Kibar Z, et al. Connexin46 mutations in autosomal dominant congenital cataract. *Am J Hum Genet.* 1999;64:1357-1364.
45. Shiels A, Mackay D, Ionides A, et al. A missense mutation in the human connexin50 gene (GJA8) underlies autosomal dominant "zonular pulverulent" cataract, on chromosome 1q. *Am J Hum Genet.* 1998;62:526-532.
46. Bu L, Jin Y, Shi Y, et al. Mutant DNA-binding domain of HSF4 is associated with autosomal dominant lamellar and Marner cataract. *Nat Genet.* 2002;31:276-278.
47. Pras E, Levy-Nissenbaum E, Bakhan T, et al. A missense mutation in the LIM2 gene is associated with autosomal recessive presenile cataract in an inbred Iraqi Jewish family. *Am J Hum Genet.* 2002;70:1363-1367.
48. Jamieson RV, Perveen R, Kerr B, et al. Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. *Hum Mol Genet.* 2002;11:33-42.
49. Berry V, Francis P, Kaushal S, Moore A, Bhattacharya S. Missense mutations in MIP underlie autosomal dominant "polymorphic" and lamellar cataracts linked to 12q. *Nat Genet.* 2000;25:15-17.
50. Khan K, Rudkin A, Parry DA, et al. Homozygous mutations in PXDN cause congenital cataract, corneal opacity, and developmental glaucoma. *Am J Hum Genet.* 2011;89:464-473.
51. Eiberg H, Lund AM, Warburg M, Rosenberg T. Assignment of congenital cataract Volkmann type (CCV) to chromosome 1p36. *Hum Genet.* 1995;96:33-38.
52. Butt T, Yao W, Kaul H, et al. Localization of autosomal recessive congenital cataracts in consanguineous Pakistani families to a new locus on chromosome 1p. *Mol Vis.* 2007;13:1635-1640.
53. Wang L, Lin H, Shen Y, et al. A new locus for inherited nuclear cataract mapped to the long arm of chromosome 1. *Mol Vis.* 2007;13:1357-1362.
54. Gao L, Qin W, Cui H, et al. A novel locus of coralliform cataract mapped to chromosome 2p24-pter. *J. Hum Genet.* 2005;50;305-310.
55. Khaliq S, Hameed A, Ismail M, Anwar K, Mehdi SQ. A novel locus for autosomal dominant nuclear cataract mapped to chromosome 2p12 in a Pakistani family. *Invest Ophthalmol Vis Sci.* 2002;43:2083-2087.
56. Abouzeid H, Meire FM, Osman I, et al. A new locus for congenital cataract, microcornea, microphthalmia, and atypical iris coloboma maps to chromosome 2. *Ophthalmology* 2009;116:154-162.

57. Sabir N, Riazuddin SA, Butt T, et al. Mapping of a new locus associated with autosomal recessive congenital cataract to chromosome 3q. *Mol Vis.* 2010;16:2634-2638.
58. Kaul H, Riazudddin SA, Yasmeen A, et al. A new locus for autosomal recessive congenital cataract identified in a Pakistani family. *Mol Vis.* 2010;16:240-245.
59. Sabir N, Riazuddin SA, Kaul H, et al. Mapping of a novel locus associated with autosomal recessive congenital cataract to chromosome 8p. *Mol. Vis.* 2010;16:2911-2915.
60. Dash DP, Silvestri G, Hughes AE. Fine mapping of the keratoconus with cataract locus on chromosome 15q and candidate gene analysis. *Mol Vis.* 2006;12:499-505.
61. Berry V, Ionides AC, Moore AT, et al. A locus for autosomal dominant anterior polar cataract on chromosome 17p. *Hum Mol Genet.* 1996;5:415-419.
62. Armitage MM, Kivlin JD, Ferrell RE. A progressive early onset cataract gene maps to human chromosome 17q24. *Nat Genet.* 1995;9:37-40.
63. Zhao R, Yang Y, He X, et al. An autosomal dominant cataract locus mapped to 19q13-qter in a Chinese family. *Mol Vis.* 2011;17:265-269.
64. Zhang S, Liu M, Dong JM, et al. Identification of a genetic locus for autosomal dominant infantile cataract on chromosome 20p12.1-p11.23 in a Chinese family. *Mol Vis.* 2008;14:1893-1897.
65. Craig JE, Friend KL, Gecz J, et al. A novel locus for X-linked congenital cataract on Xq24. *Mol Vis.* 2008;14:721-726.

FIGURE LEGENDS

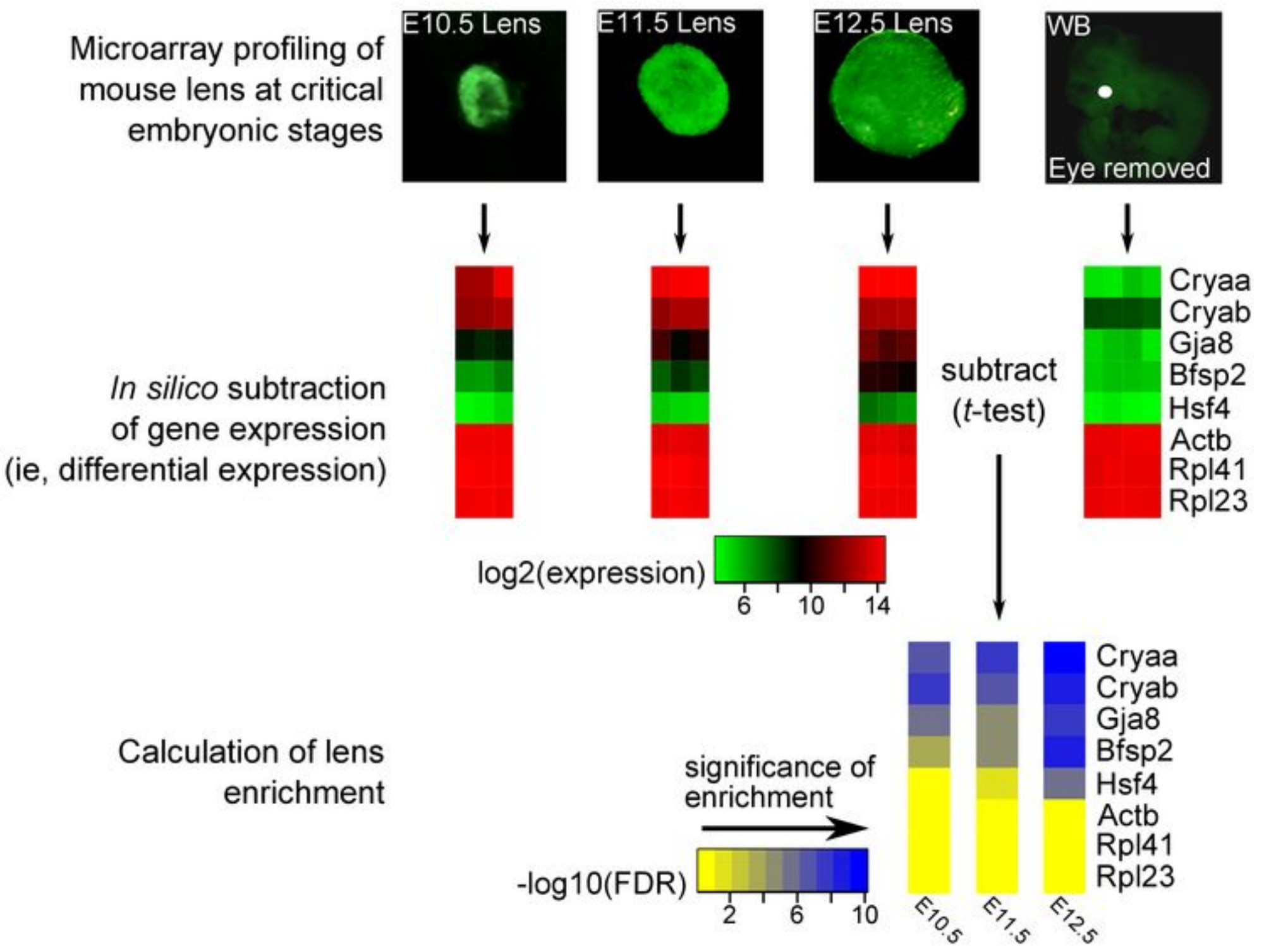
Figure 1. Strategy for building *iSyTE*. To identify genes that are specifically expressed in the lens during embryonic development, mouse embryonic lens tissue at E10.5, E11.5, and E12.5 was profiled using microarrays. Several hundred lenses at stages E10.5, E11.5 and E12.5 were pooled for generating total RNA for each biological replicate in microarrays conducted in triplicate. We also obtained the microarray gene expression profile for pooled whole embryonic body (WB) tissue at each stage; the ocular region was removed from the whole embryonic tissue before profiling. Lens-specific profiles are "subtracted" from the WB control using a moderated *t*-test. A lens enrichment *p*-value is assigned to each gene for each embryonic stage, and a False Discovery Rate (FDR) was calculated based on the *p*-value. *t* statistics were used to rank the genes for lens enrichment. The green color of the lenses represents fluorescence due to use of lens tissue carrying a *Pax6P03.9-GFP* reporter, included in pilot experiments as a quality control for the fidelity of lens collection (see Methods).

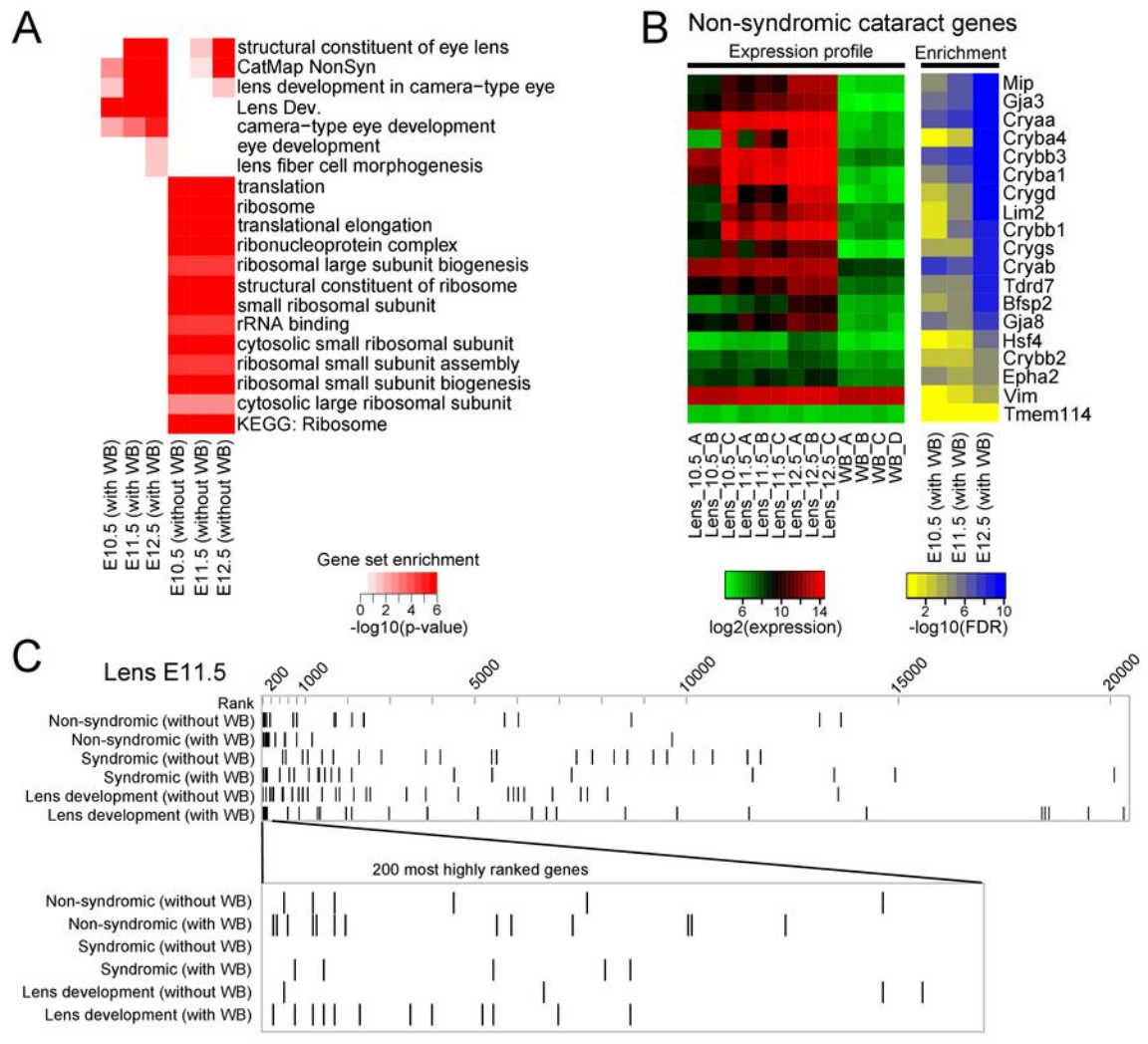
Figure 2. *In silico* subtraction is an effective tool to identify lens-enriched genes. (A) The 200 most highly ranked genes with WB subtraction and without WB subtraction (No WB) at E10.5, E11.5, and E12.5 were tested against many functional biological gene categories to identify statistically significantly enriched gene sets (Fisher's exact test, Bonferroni corrected $p<0.05$, odds ratio of gene set overlap > 20). Significantly enriched genes sets are visualized in the heat map. **(B)** Heat maps representing expression levels and lens enrichment *p*-values of all non-syndromic human cataract genes cataloged at CatMap. **(C)** A rank list showing the distribution of known genes related to human cataract and embryonic lens development based on the lens enrichment *t*-statistics (with WB) or microarray expression (without WB). The ranked list of the 200 most highly ranked genes is expanded and shown underneath the full ranked list.

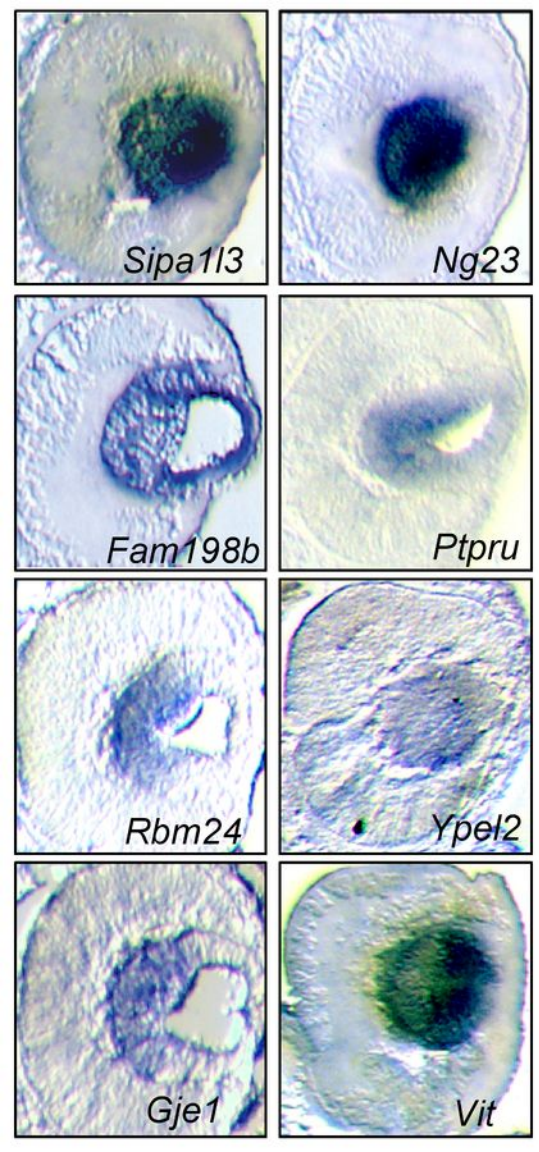
Figure 3. *iSyTE* predicts potential candidate genes in mapped cataract loci in human and mouse. Section *in situ* hybridization on E11.5-12.0 mouse embryonic tissue confirms lens expression for *Sipa1l3* (Human locus 19q13.13, *SIPA1L3*), *Ptpru* (Human locus 1p35.3, *PTPRU*), *Ng23* (Human locus 6p21.33, *C6orf26*), *Fam198b* (Human locus 4q32.1, *FAM198B*), *Rbm24* (Human locus, 6p22.3, *RBM24*) *Ypel2* (Human locus 17q22, *YPEL2*), *Gje1* (Human locus 6q24.1, *GJE1*) and *Vit* (Human locus, 2p22.2, *VIT*).

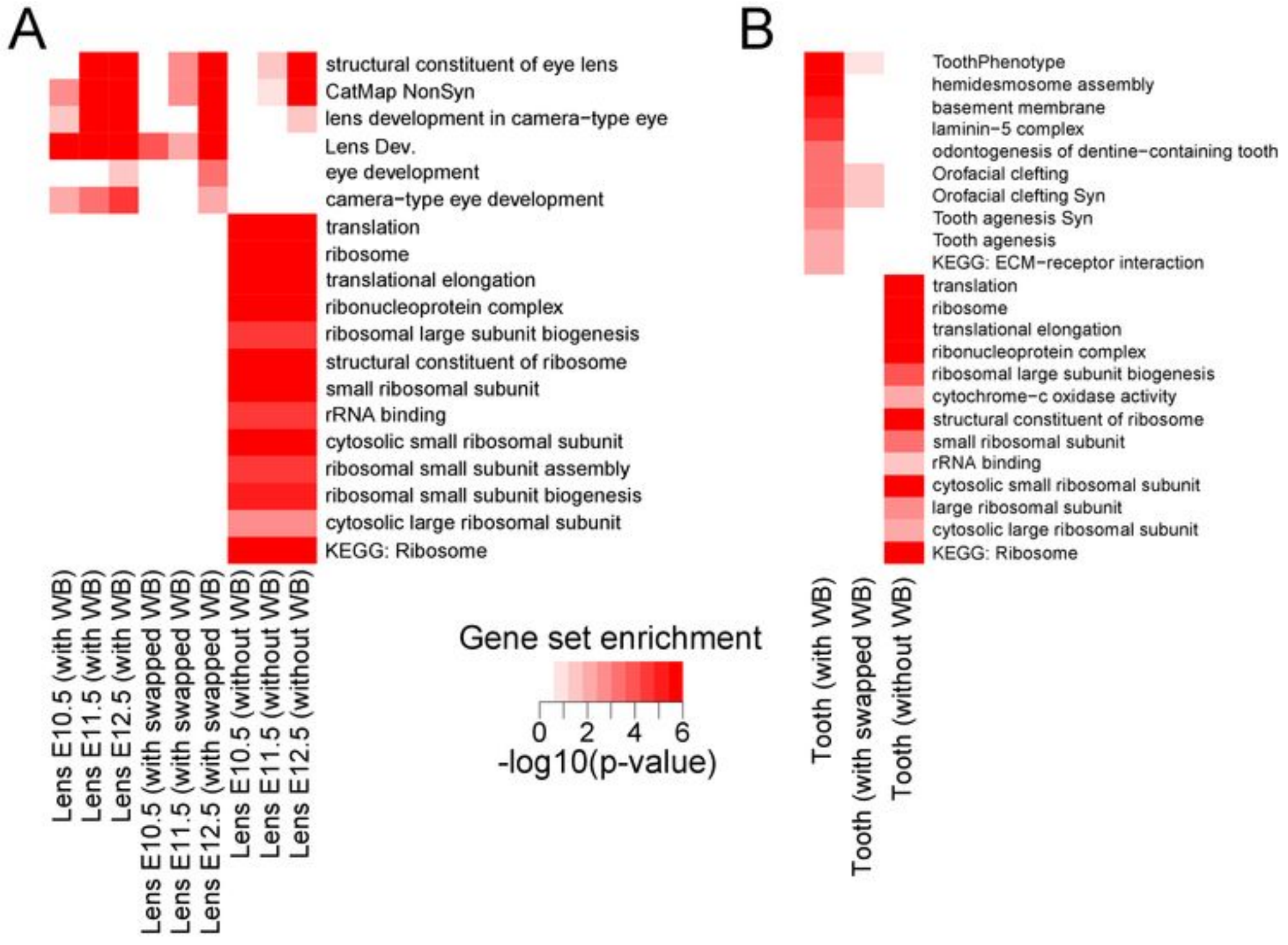
Figure 4. *In silico* subtraction strategy is robust against use of different WB controls. After swapping WB control profiles generated for separate lens and tooth analyses, the *in silico* subtraction strategy still robustly identifies genes that are specific to: **(A)** lens and **(B)** tooth. Thus, the *in silico* subtraction strategy is robust against the use of different WB. This supports the idea that the WB generated in this study can be used as a public resource for comparison with gene expression profiles of other embryonic tissues at similar stages.

Figure 5. *iSyTE* tracks on the UCSC Genome Browser represent a publically available resource for cataract gene identification. *iSyTE* custom tracks (accessible at http://bioinformatics.udel.edu/Research/iSyTE) visualize the ranking of the lens enrichment for E10.5, E11.5, and E12.5 mouse lens. *iSyTE* tracks on the UCSC Genome Browser identify genes associated with primary congenital cataract in human and mouse. This figure visualizes a 10 Mb locus on human chromosome 9 with *iSyTE* tracks, which confirms a non-syndromic cataract associated gene, *TDRD7*, as the most promising candidate in this interval. Color of the *iSyTE* tracks is based on the rank of the individual genes after WB subtraction.


Microarray profiling of mouse lens at critical embryonic stages
E10.5 Lens
E11.5 Lens
E12.5 Lens
WB
Eye removed
In silico subtraction of gene expression (ie, differential expression)
subtract (t-test)
Cryaa
Cryab
Gja8
Bfsp2
Hsf4
Actb
Rpl41
Rpl23
log2(expression)
6
10
14
Calculation of lens enrichment
significance of enrichment
-log10(FDR)
2
6
10
E10.5
E11.5
E12.5


A
structural constituent of eye lens
CatMap NonSyn
lens development in camera−type eye
Lens Dev.
camera−type eye development
eye development
lens fiber cell morphogenesis
translation
ribosome
translational elongation
ribonucleoprotein complex
ribosomal large subunit biogenesis
structural constituent of ribosome
small ribosomal subunit
rRNA binding
cytosolic small ribosomal subunit
ribosomal small subunit assembly
ribosomal small subunit biogenesis
cytosolic large ribosomal subunit
KEGG: Ribosome
E10.5 (with WB)
E11.5 (with WB)
E12.5 (with WB)
E10.5 (without WB)
E11.5 (without WB)
E12.5 (without WB)
Gene set enrichment
0 2 4 6
-log10(p-value)
B
Non-syndromic cataract genes
Expression profile
Enrichment
Mip
Gja3
Cryaa
Cryba4
Crybb3
Cryba1
Crygd
Lim2
Crybb1
Crygs
Cryab
Tdrd7
Bfsp2
Gja8
Hsf4
Crybb2
Epha2
Vim
Tmem114
Lens_10.5_A
Lens_10.5_B
Lens_10.5_C
Lens_11.5_A
Lens_11.5_B
Lens_11.5_C
Lens_12.5_A
Lens_12.5_B
Lens_12.5_C
WB_A
WB_B
WB_C
WB_D
E10.5 (with WB)
E11.5 (with WB)
E12.5 (with WB)
6 10 14
log2(expression)
2 6 10
-log10(FDR)
C
Lens E11.5
200
1000
5000
10000
15000
20000
Rank
Non-syndromic (without WB)
Non-syndromic (with WB)
Syndromic (without WB)
Syndromic (with WB)
Lens development (without WB)
Lens development (with WB)
200 most highly ranked genes
Non-syndromic (without WB)
Non-syndromic (with WB)
Syndromic (without WB)
Syndromic (with WB)
Lens development (without WB)
Lens development (with WB)


Sipa1l3
Ng23
Fam198b
Ptpru
Rbm24
Ypel2
Gje1
Vit


A
structural constituent of eye lens
CatMap NonSyn
lens development in camera−type eye
Lens Dev.
eye development
camera−type eye development
translation
ribosome
translational elongation
ribonucleoprotein complex
ribosomal large subunit biogenesis
structural constituent of ribosome
small ribosomal subunit
rRNA binding
cytosolic small ribosomal subunit
ribosomal small subunit assembly
ribosomal small subunit biogenesis
cytosolic large ribosomal subunit
KEGG: Ribosome
Lens E10.5 (with WB)
Lens E11.5 (with WB)
Lens E12.5 (with WB)
Lens E10.5 (with swapped WB)
Lens E11.5 (with swapped WB)
Lens E12.5 (with swapped WB)
Lens E10.5 (without WB)
Lens E11.5 (without WB)
Lens E12.5 (without WB)
Gene set enrichment
0 2 4 6
-log10(p-value)
B
ToothPhenotype
hemidesmosome assembly
basement membrane
laminin−5 complex
odontogenesis of dentine−containing tooth
Orofacial clefting
Orofacial clefting Syn
Tooth agenesis Syn
Tooth agenesis
KEGG: ECM−receptor interaction
translation
ribosome
translational elongation
ribonucleoprotein complex
ribosomal large subunit biogenesis
cytochrome−c oxidase activity
structural constituent of ribosome
small ribosomal subunit
rRNA binding
cytosolic small ribosomal subunit
large ribosomal subunit
cytosolic large ribosomal subunit
KEGG: Ribosome
Tooth (with WB)
Tooth (with swapped WB)
Tooth (without WB)

UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr9:94,000,000-104,000,000 gene jump clear size 10,000,001 bp. configure

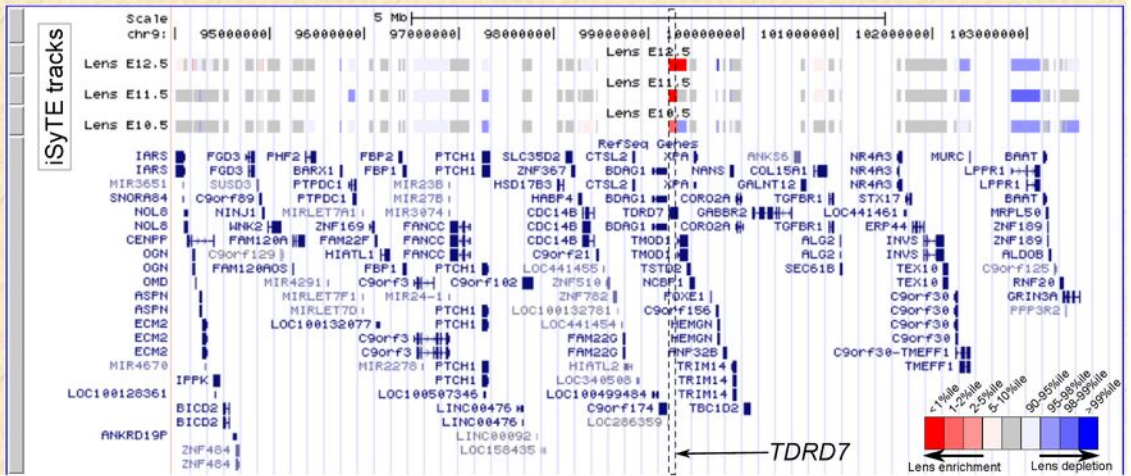


Table 1. *iSyTE* rank of genes associated with human isolated congenital cataract.

Gene	Chr	Interval Size (Mb)	No. Genes	*iSyTE* Rank	Reference
BFSP1	20	5.43	22	1	29
BFSP2	3	17.88	106	1	30
CHMP4B	20	3.03	43	34	31
CRYAA1	21	2.79	37	1	32
CRYAB	11	21.08	99	1	33
CRYBA1	17	15.20	129	1	34
CRYBA4	22	0.40	6	1	35
CRYBB1	22	2.54	21	2	36
CRYBB2	22	1.47	5	2	37
CRYBB3	22	3.76	31	1	38
CRYGC	2	32.97	129	1	39
CRYGD	2	32.97	129	2	39
CRYGS	3	2.76	31	1	40
EPHA2	1	5.75	67	1	41
FYCO1	3	12.21	191	21	42
GCNT2	6	5.26	21	7	43
GJA3	13	8.69	50	1	44
GJA8	1	46.04	299	1	45
HSF4	16	11.41	102	2	46
**LIM2*	19	6.71	103	1	47
MAF	16	5.01	18	1	48
MIP	12	24.00	144	1	49
PXDN	2	6.68	18	1	50
TDRD7	9	21.08	108	1	16

iSyTE rank is obtained by comparing the lens enrichment of all the genes within a given interval at each of E10.5, E11.5, and E12.5, and the overall *iSyTE* rank is the minimum of the ranking of these three stages.

* Mutations in *LIM2* are associated with both congenital and percentile cataract.

Table 2. Signal intensities for gene expression in lens and WB.

Gene	E10.5 Lens	E11.5 Lens	E12.5 Lens	WB	Gene Name
Cryaa	8511	18154	21441	32	Crystallin, alpha A
Cryab	5334	6123	6910	229	Crystallin, alpha B
Cryba1	4692	14893	22417	26	Crystallin, beta A1
Cryba4	280	840	18057	35	Crystallin, beta A4
Crybb1	1420	9050	19732	63	Crystallin, beta B1
Crybb2	135	108	171	50	Crystallin, beta B2
Crybb3	8032	13934	23115	83	Crystallin, beta B4
Crygb	140	161	8990	12	Crystallin, gamma B
Crygc	6875	8054	20487	37	Crystallin, gamma C
Crygd	983	729	13536	23	Crystallin, gamma D
Crygs	547	734	2570	17	Crystallin, gamma S
Gja3	700	1984	4876	15	Gap junction protein, alpha 3
Gja8	429	1011	2351	39	Gap junction protein, alpha 8
Bfsp1	147	114	707	27	Beaded filament structural protein 1, filensin
Bfsp2	90	233	985	43	Beaded filament structural protein 2, phakinin
Lim2	537	2149	8179	75	Lens intrinsic membrane protein 2
Mip	525	1008	6613	25	Major intrinsic protein of eye lens fiber
Epha2	249	255	295	80	Eph receptor A2
Foxe3	1111	1588	1033	61	Forkhead box E3
Hsf4	27	35	106	21	Heat shock transcription factor 4
Maf	1263	1228	1465	96	v-maf musculoaponeurotic fibrosarcoma (avian) oncogene homolog
Pitx3	1279	1743	1429	77	Paired-like homeodomain transcription factor 3
Meis1	608	373	301	587	Meis homeobox 1
Bmp7	214	147	105	115	Bone morphogenetic protein 7
Pax6	3838	2980	2500	164	Paired box gene 6
Sox2	1670	496	244	963	SRY-box containing gene 2
Sox1	504	917	1314	65	SRY-box containing gene 1
Six3	1386	1121	837	74	SIX homeobox 3
Mab21l1	2398	1914	1596	228	Mab-21-like 1 (*C. elegans*)
Prox1	502	684	717	20	Prospero-related homeobox 1
Tdrd7	866	1037	4096	108	Tudor domain containing 7

Table 3. *iSyTE* predicted candidate genes in mapped intervals for human cataract.

Chr Location (Reference)	No. genes (minRank ≤ 500)* / No. genes in interval	Genes with minRank ≤ 500 (in order)	Ref.
1pter-p36	3 / 36	*MXRA8, FAM132A, C1orf159*	51
1p35.3-p36.32	18 / 266	*LIN28A, PTPRU, ALPL, MAN1C1, AGTRAP, DNAJC16, EPHA2, ESPN, C1QB, CD52, SLC25A33, C1QA, MECR, SPSB1, KLHL21, NIPAL3, FAM54B, CLCN6*	19
1p34.3-p32.2	11 / 199	*DMRTA2, FOXE3, RSPO1, SLC2A1, PTPRF, YBX1, HYI, PRKAA2, AKIRIN1, INPP5B, C1orf109*	52
1q25-q31	0 / 1	-	53
2p24-pter	1 / 39	*PXDN*	54
2p12	2 / 11	*TACR1, FAM176A*	55
2q33	0 / 26	-	56
3q26.1-3q27.2	8 / 84	*CRYGS, GPR160, PLD1, PRKCI, ETV5, TTC14, FAM131A, MAP6D1*	57
7q21.11-q31.1	11 / 179	*NRCAM, STEAP1, PON2, C7orf51, SEMA3A, LAMB1, CDK6, C7orf23, SLC25A13, STEAP2, SLC25A40*	58
8p23.2-p21.3	5 / 78	*SLC7A2, MSRA, RHOBTB2, MTMR7, FAM86B2*	59
15q22.32-q24.2	3 / 45	*TMED3, PEAK1, FAH*	60
17p13	1 / 90	*ENO3*	61
17p24	0 / 15	-	62
19q13	8 / 201	*SIPA1L3, PRX, EML2, SPINT2, PVRL2, PLD3, SLC1A5, TRAPPC6A*	18
19q13-qter	1 / 88	*LENG8*	63
20p11.23-p12.1	1 / 29	*BFSP1*	64
Xq24	3 / 35	*NDUFA1, UPF3B, AKAP14*	65

*minRank is the smallest rank value of that gene across the three embryonic stages with respect to all the genes in the lens microarray. For example, if a gene is ranked 12, 401, and 214 at E10.5, E11.5 and E12.5 respectively, the minRank is 12. A gene with minRank below 500 will have a bright red color in the *iSyTE* track in UCSC Genome Browser.