DELocal: Chromosomal neighbourhoods having genes of diverse functions allow improved identification of differentially expressed genes

Rishi Das Roy[1*], Outi Hallikas[1], Mona M. Christensen[1], Elodie Renvoisé[1,2], Jukka Jernvall[1,3*]

[1]Institute of Biotechnology, University of Helsinki, P.O. Box 56, FI-00014 Helsinki, Finland.
[2]5 Rue de Portland, 72100 Le Mans, France.
[3]Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, FI-00014 Helsinki, Finland.

[*]To whom correspondence should be addressed. Email: rishi.dasroy@helsinki.fi and jernvall@fastmail.fm

ABSTRACT

Exploration of genetically modified organisms, developmental processes, diseases or responses to various treatments require accurate measurement of changes in gene expression. This can be done for thousands of genes using high throughput technologies such as microarray and RNAseq. However, identification of differentially expressed (DE) genes poses technical challenges due to limited sample size, few replicates, or simply very small changes in expression levels. Consequently, several methods have been developed to determine DE genes, such as Limma, RankProd, SAM, and DeSeq2. These methods identify DE genes based on the expression levels alone. As genomic co-localization of genes is generally not linked to co-expression, we deduced that DE genes could be detected with the help of genes from chromosomal neighbourhood. Here, we present a new method, DELocal, which identifies DE genes by comparing their expression changes to changes in adjacent genes in their chromosomal regions. Our results show that DELocal provides distinct benefits in the identification of DE genes. Furthermore, our comparative analysis of the dispersal of genes with related functions suggests that DELocal is applicable to a wide range of developmental systems. With increasing availability of genomic data, gene neighbourhood can become a powerful tool to detect differential expression.

INTRODUCTION

One key aspect of development and differentiation is regulation of gene expression. During development, genes are expressed highly dynamically, and perturbations of the expression dynamics underlie many diseases and developmental defects. Developmental regulation of gene expression through time can be examined by quantifying gene expression levels at two or more time points. In general, expression dynamics of genes under the same regulation, or expression of genes involved in the same biological function can be expected to correlate to some extent. In fact, this is commonly the case in prokaryotes, in which co-functioning genes are expressed under a single promoter in a shared operon (1). In eukaryotes, it is not completely understood how thousands of genes are precisely regulated (2). Gene expression is initiated by transcription factors interacting with enhancers which are usually located at different distances from the target genes. In vertebrates, there are only few large gene clusters, such as Hox genes and immunoglobulin genes, whose spatial organization on the chromosome is crucial to their regulation and function. This kind of concerted expression of adjacent genes has generally thought to have evolved via tandem duplications of ancestral genes. In some cases co-regulated or co-expressed genes have been reported to be also co-localized within the same genomic regions and chromosomes, or be nearby to each other in three-dimensional space due to folding of the DNA (3-5). Nevertheless, a strict spatial co-localization of co-regulated genes is not the general pattern (6), and neighbouring genes may show highly distinct expression dynamics.

In this paper we first investigate the differential expression and the spatial distribution of genes regulating a specific organ system, the mammalian tooth. Genes required for normal progression of tooth development are well characterized (7) and references therein), and the developmental regulation of teeth shares many similarities with other epithelial organs, such as feathers and hair. Our aim is to test whether inclusion of gene neighbourhood in the analyses provides additional benefits in detection of differential expression. For the analyses, we developed an algorithm, called DELocal to identify differentially expressed (DE) genes based on their neighbours' expression patterns. Using embryonic mouse dental transcriptomes obtained with both microarray and RNAseq, we show how DELocal compares favourably to other methods to identify DE genes (8-13). Finally, we show that genes sharing the same gene ontology terms are dispersed in the chromosomal regions of mice and humans, suggesting the general potential for using gene neighbourhood to detect differential expression.

MATERIALS AND METHODS

Ethics statement

All mouse studies were approved and carried out in accordance with the guidelines of the Finnish national animal experimentation board under licenses KEK16-021, ESAVI/2984/04.10.07/2014 and ESAV/2363/04.10.07/2017.

RNAseq library preparation

Developing mouse molar teeth from embryonic days 13.5 (E13) and 14.5 (E14) were dissected from wild type C57BL/Ola embryos. For RNAseq, five biological replicates were used. The samples were stored in RNAlater (Qiagen GmbH, Hilden, Germany) in -75°C. RNA was extracted first twice with guanidinium thiocyanate-phenol-chloroform extraction and then further purified using RNeasy Plus micro kit (Qiagen GmbH, Hilden, Germany) according to manufacturer's instructions. RNA quality of representative samples was assessed with 2100 Bioanalyzer (Agilent, Santa Clara, CA) and the RIN values were 9 or higher. The RNA concentrations were determined by Qubit RNA HS Assay kit (Thermo Fisher Scientific, Waltham, MA). The cDNA libraries were prepared with Ovation RNAseq System V2 (Nugene, Irvine, CA), and sequenced with NextSeq500 (Illumina, San Diego, CA).

Microarray library preparation

Mouse E13 and E14 teeth were dissected from wild type NMRI embryos. Five biological replicates were used. The amount of RNA available in each sample was measured with 2100 Bioanalyzer (Agilent, Santa Clara, CA). Only the samples showing a RIN value above 9 were used for the microarray analysis.

Gene Expression analysis

Gene expression was measured both in microarray (platform: GPL6096, Affymetrix Mouse Exon Array 1.0) and RNAseq (platforms GPL19057, Illumina NextSeq 500). The microarray gene signals were normalized with aroma.affymetrix (14) package using Brainarray custom CDF (Version 19, released on Nov 13, 2014) (15). The RNAseq reads (84 bp) were evaluated and bad reads were filtered out using FastQC (16), AfterQC (17) and Trimmomatic (18). This resulted in on average 63 million reads per sample. Then good reads were aligned with STAR (19) to Mus_musculus.GRCm38.dna.primary_assembly.fa genome and counts for each gene was performed by HTSeq (20) tool using Mus_musculus.GRCm38.90.gtf annotation. On average 89% reads were uniquely mapped to Mus musculus genome. Additionally, RNAseq count values were normalized using DESeq2 (13). All the transcriptomic data are available in GEO under the accession number GSE142201.

DELocal

In DELocal, it is hypothesized that differentially expressed genes have different expression dynamics compared to their neighbouring genes. We used a similar logic to ESLiM (21), an algorithm which detects changes in exon usage. In an analogy, a neighbourhood could be assumed as a single gene and genes in the neighbourhood are equivalent to exons of the gene. Then identifying a DE gene is comparable to detecting alternative splicing.

In this algorithm, gene's expression is modelled as a linear relationship with median expression of neighbourhood genes, such as,

$$\hat{g}_{ij} = s_i \times N\widetilde{g}w_{ij} + b_i \qquad \text{....... (i),}$$

where $\hat{g}_{ij}$ is expected expression of *i*-th gene in *j*-th sample, $N\widehat{gw}_{ij}$ is median expression of *N* nearest neighbouring genes within 1 Mb window of the *i*-th gene from *j*-th sample and b_i is base line expression level of gene g_i. The slope s_i of every gene g_i depends on its neighbouring genes. Therefore, the difference between expected and observed values or residual,

$r_{ij} = g_{ij} - \hat{g}_{ij}$ …… (ii)

where g_{ij} is observed value. For DE genes, these residual values will be significantly different in different biological conditions.

Furthermore, with the aid of the residuals r_{ij} observed g_{ij} could be formulated as follow,

$g_{ij} = s_i \times \widetilde{Ng}w_{ij} + b_i + r_{ij}$ …. (iii)

Noticeably, this relationship Eq. (iii) is independent of experimental condition and only dependent on neighbouring gene. Therefore, similar to ESLiM, DE genes are detected through significantly deviated residual values between the desired contrasts using Empirical Bayes statistics, available from Limma package (11). We tested the performance of DELocal using from 1 to 14 neighbouring genes (N in Eq. (iii)). The log-normalized and normalized count values were used in DELocal respectively for microarray and RNAseq data. There are 334 protein coding genes in mouse genome which do not have any other protein coding gene in their 1Mb neighbourhood. Therefore, we also used available non-protein coding genes from the neighbourhood in DELocal analysis. However, after the inclusion of non-protein coding genes there are still 17 protein coding genes without any neighbours within 1 Mb.

Performance measures

DELocal was compared with different publicly available tools applicable both for microarray or RNAseq : RankProd(8), SAM(9), DEMI(10), Limma(11), edgeR(12) and DESeq2 (13). All these programs were executed with default parameters. Genes reported with p-value <= 0.05 by these tools were marked as differentially expressed gene and used to evaluate the performance of each tool using the following metrices and receiver operating characteristic (ROC) curves.

- Sensitivity (Recall) - TPR, true positive rate TPR = TP / (TP + FN)
- Specificity - SPC, true negative rate SPC = TN / (TN + FP)
- Precision - PPV, positive predictive value PPV = TP / (TP + FP)
- Accuracy - ACC = (TP + TN) / (TP + FP + FN + TN)
- F1 score F1 = 2TP / (2TP + FP + FN)
- Mathews correlation coefficient (MCC) = $\frac{TP \times TN - FP \times FN}{\sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}}$

where TP, true positive; FP, false positive; TN, true negative; FN, false negative

302 genes linked to tooth development were used to find the true and false positive rate for the analyses (Supplementary Table S1 and refs (7,22)). The areas under the ROC curves were calculated with ROCR (23). The biotypes and chromosome locations of genes are downloaded from ensemble

Biomart using R script (24). Throughout the study starting locations of the genes are used as the position of the gene in the chromosome and used to measure the distances.

RESULTS AND DISCUSSION

As our focus of interest is developmental regulation, we first obtained an overall pattern of distribution of developmental genes by tabulating how closely genes are located in the genome. For example, the human genome (GRCh38.p13) is 4.5*10-9 base pairs long and has 20,449 protein coding genes, which means, on average, one gene for every 220,060 bases (Ensembl release 99). Similarly, for the mouse genome, this number is 154,290 (GRCm38.p6). Consequently, on average, 5 to 6 genes should reside in each 1 Mb window in the mouse genome. To express these statistics as a neighbourhood, we can state that each gene has 4 to 5 neighbouring genes within a 1Mb window. This observation is obviously a broad generalization, but it does indicate that genes tend to have some neighbours within 1 Mb. To examine the neighbourhood patterns in more detail, we examined the 1Mb neighbourhoods of protein coding genes of the mouse genome. For every gene, number of neighbouring genes within 1 Mb window was counted (Figure 1A, Supplementary Figure S1). This simple calculation shows that majority of protein coding genes in mouse have more than 4 neighbours, the median number of neighbours being 15 (Supplementary Figure S2). This tabulation indicates that there is some level of clustering of genes in the eukaryotic genome, a pattern well established in the literature (25).

Next, we examined genes associated with the development of the mouse tooth. This single organ focus allowed us to test whether genes participating in the regulation of the same organ are also located close to each other. Here mouse molar development provides a good example because its gene regulation is relatively well understood, and because tooth development itself is relatively autonomous process (7,26). We analysed the 1 Mb neighbourhoods of the tooth developmental genes (TDG, Supplementary Table S1) and the number of TDGs that are sharing the same neighbourhood. The results show that TDGs are mostly located far from each other, suggesting that genes regulating this specific developmental process are not co-localized (Figure 1B). Taken together, although majority of protein coding genes are to some extent clustered in the mouse genome, TDGs tend to be located far from each other (Figure 1).

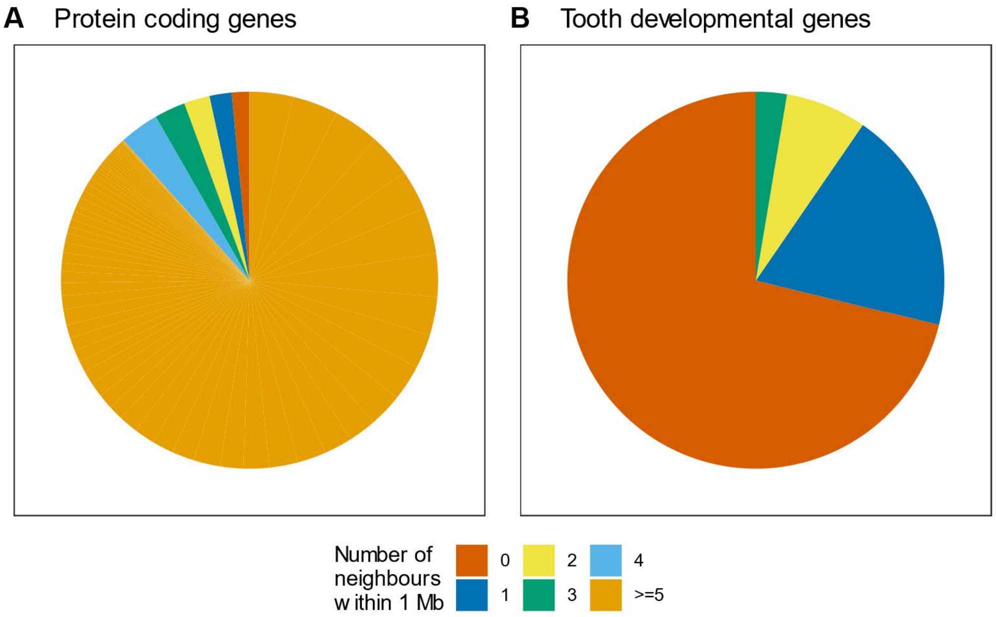

Figure 1. Although protein-coding genes typically have neighbours, tooth genes lack other tooth gene neighbours within 1 Mb windows around each gene. (**A**) The number of neighbouring genes within 1 Mb window around each gene tabulated from the mouse genome for all protein coding genes and (**B**) genes involved in tooth development. Majority of 21,971 protein coding genes have at least five neighbours whereas most of 302 tooth developmental genes lack tooth genes as neighbours. This pattern suggests that genes with specific functions are sparsely distributed.

DELocal method to detect differentially expressed genes

Because genes linked to tooth development do not appear to be co-localized in the genome, we decided to examine whether DE genes could be identified by comparing their level of gene expression with their neighbouring genes. We examined differential expression of genes at the onset of tooth crown formation, between embryonic day 13.5 (E13) and 14.5 (E14) when many of the TDGs are known to be upregulated (7,27). For example, *Ctnna1*, *Shh*, *Foxi3* and *Sostdc1*, all genes required for normal tooth developmental (28), show prominent upregulation between E13 and E14 compared to the other genes in their neighbourhoods (Figure 2). Building from this observation, next we developed a new algorithm, DELocal, to identify DE genes based on their neighbours' expression dynamics. To evaluate its performance, we used gene expression data from embryonic mouse dental tissues generated by both microarray and RNAseq.

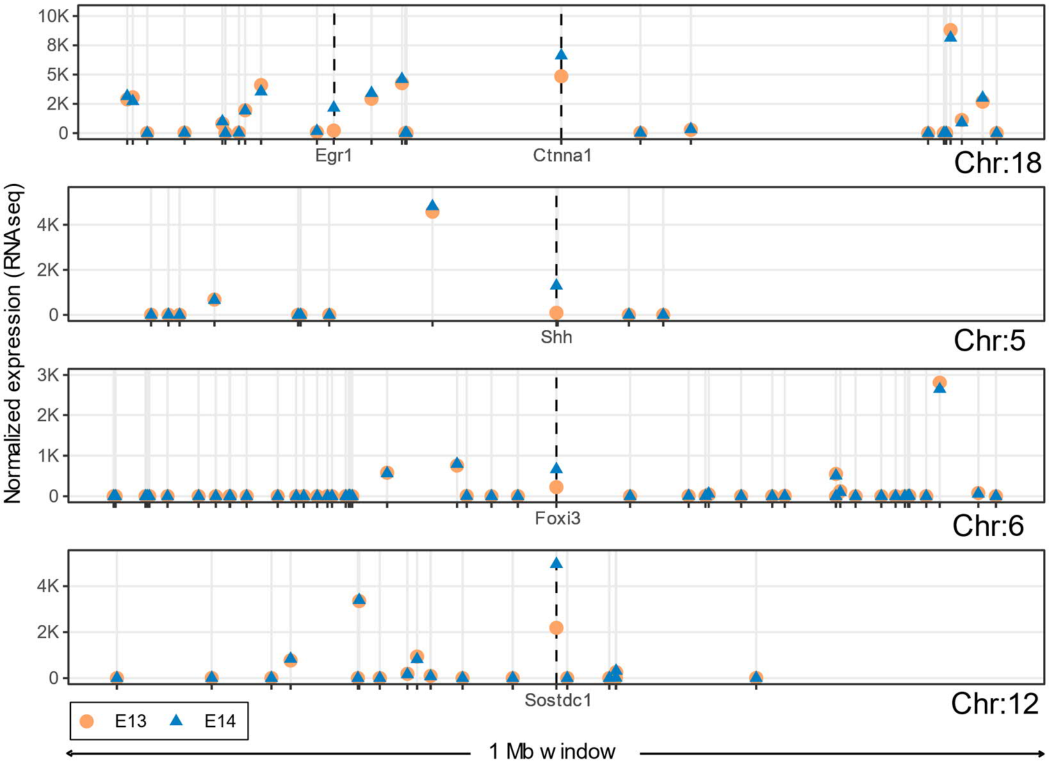

Figure 2. Only tooth developmental genes are differentially expressed within 1 Mb windows in developing mouse molar. Median expression levels of the tooth developmental genes *Ctnna1*, *Shh, Foxi3* and *Sostdc1* and their neighbouring genes at developmental stages E13 and E14. Regardless of the expression level, the surrounding genes show little change between the two stages, compared to the tooth developmental genes. *Egr1* in the 1 Mb window of *Ctnna1* is also a tooth developmental gene.

Optimizing the number of neighbours

Our hypothesis of neighbouring genes being informative in the detection of differential expression is dependent on the definition of 'neighbourhood'. Therefore, it is important to determine the right number of neighbours to include in the analysis by the DELocal algorithm. To define the optimal number of neighbours we tested different numbers (1-14) of closest genes within a fixed window (1 Mb) surrounding the gene of interest. We evaluated the performance of DELocal with different numbers of closest neighbours in identifying the genes involved in tooth development (TDGs). Again we contrasted the expression levels between E13 to E14 molar teeth, or so-called bud stage to cap stage transition, when many TDGs are known to be active (7). The Matthews correlation coefficient (MCC) scores were measured for different numbers of neighbours to examine the strength of DELocal to identify TDGs (true positive;TP) as well as non-TDGs (true negatives; TN). The MCC score was chosen to optimize the model due to very few TPs, or imbalanced dataset. The results show that DELocal produces similar and stable MCC scores on both microarray and RNAseq datasets, even though RNAseq data produces slightly higher MCC scores than microarray (Figure 3). We note that only one nearest gene is enough to obtain close to the highest MCC score. However, for RNAseq the best MCC score corresponds to 5 nearest neighbours. Because there are fewer genes available in microarray analyses compared to RNAseq, in the rest of the study we used DELocal with 5 neighbours both for the RNAseq and microarray data.

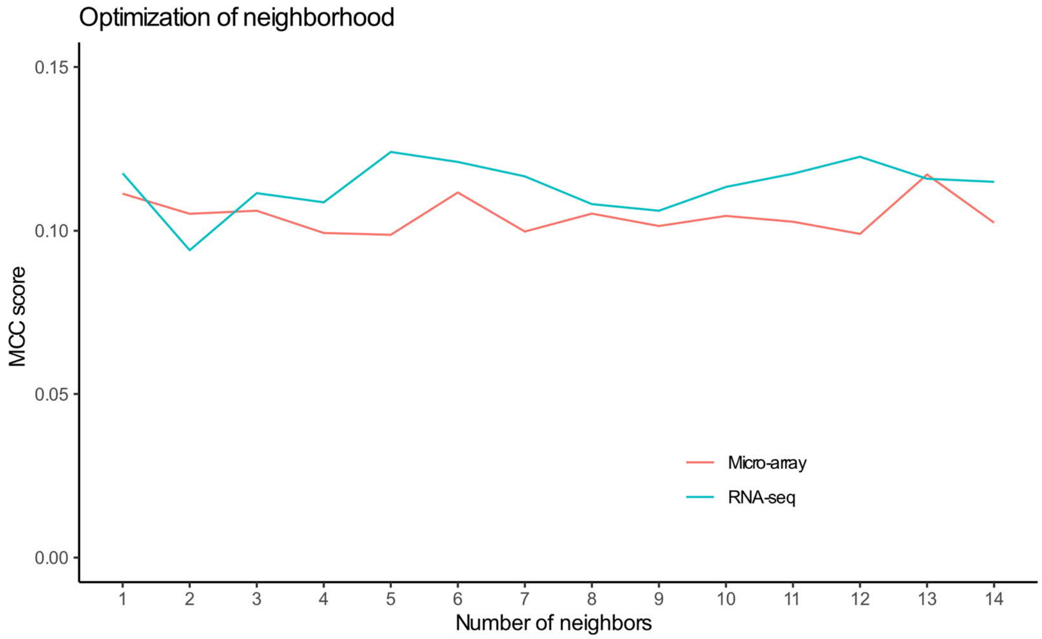

Figure 3. DELocal performance is not strongly dependent on the number of gene neighbours used in the analysis. Every gene is evaluated in relation to its neighbouring genes. In the absence of any "gold standard" for the number of neighbours, different numbers of nearest genes (within 1 Mb window) were used to identify the DE genes. The overall performances were measured using MCC. The performance of DELocal using RNAseq data was slightly better than with microarray data.

Comparison with other methods

Microarray is one of the earliest successful high throughput technologies to measure a large number of gene expressions, and consequently there are a good number of statistical methods to identify DE genes from datasets generated by this platform. Hence, the performance of DELocal can be evaluated by comparisons to these methods using a microarray dataset. However, microarray is limited to only the genes which have been targeted by microarray probes. Therefore, the expression of all the genes cannot be accessed, resulting in fewer neighbouring genes being sampled. To obtain a more comprehensive readout of DE genes, RNAseq was also used to evaluate DELocal performance. The performance of all these methods was measured by the ability to identify differentially expressed TDGs.

For analysis of performance, we used the receiver operating curve (ROC) (23) depicting the true positive rate against the false positive rate of DE genes. The analyses show that DELocal outperforms other methods in identifying TDGs using both microarray and RNAseq (Figure 4A, B). The performance is most similar to Limma/DEMI (microarray) and Limma/DESeq (RNAseq). We used also other metrices like specificity, recall (sensitivity), precision and MCC to evaluate and compare the different methods. DELocal shows high specificity and accuracy compared to other methods for microarray data (Supplementary Figure S3).

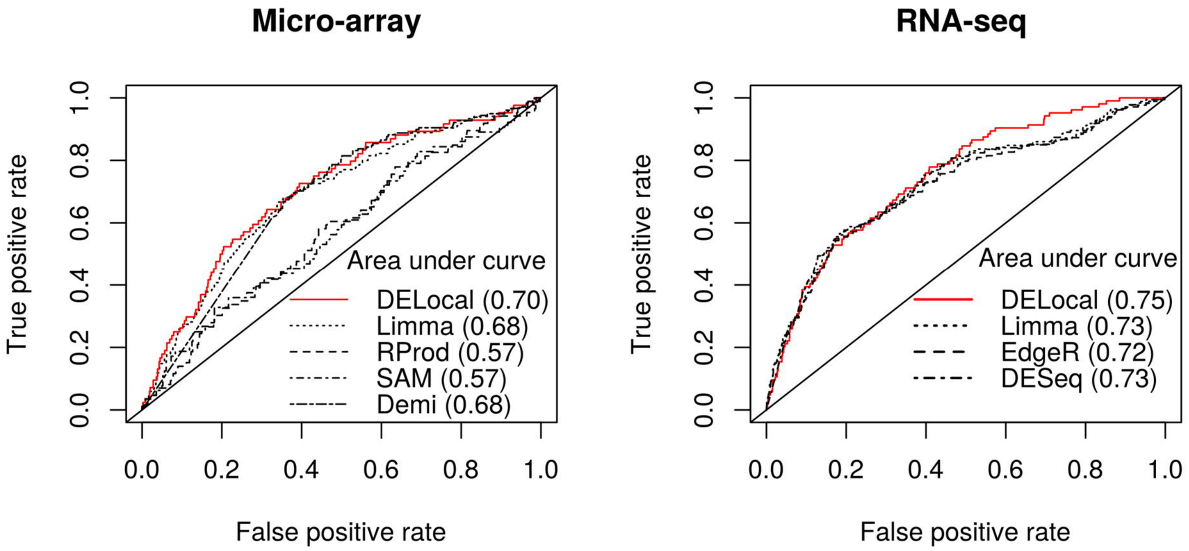

Figure 4. Compared to other methods, DELocal is powerful in detecting differentially expressed genes. Receiver operating characteristic (ROC) curves and areas under the curves (within the parenthesis) show that DELocal outperforms other methods on both microarray and RNAseq data.

In RNAseq data, DELocal outperforms other methods except in recall (sensitivity) (Figure 5). The MCC scores remain equivalent to each other. The TDG dataset is imbalanced due to the large number of non-TDGs (true negatives), which hinders the evaluation of accuracy, but does not affect F1 or MCC. Considering that the objective of many experiments is to find true positives, the F1 score, which is a compound-term of precision and recall, is an important metrics. The F1 score ranges from zero (bad) to one (good), and the F1 scores of all of the methods remain suboptimal. Nevertheless, DELocal using RNAseq clearly outperforms other methods also in F1 score. Below we discuss the results from RNAseq only.

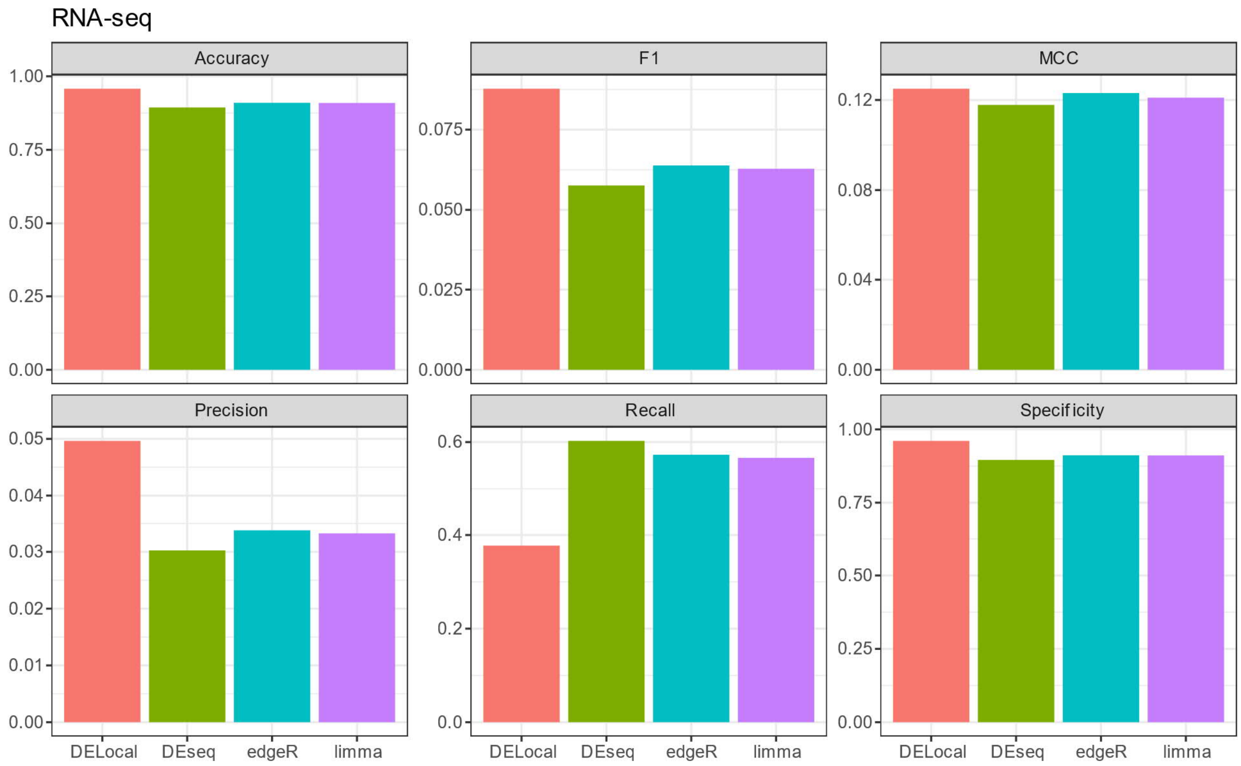

Figure 5. Comparison of DELocal with earlier methods of identifying differential expression. Evaluation matrices show that, except for recall (sensitivity), DELocal outperforms earlier methods in every matrix. However due to large number of true negatives, the significance scores of precision, F1 and MCC remained negligible. The evaluation matrices are explained in materials and method section. The analysis was done using RNAseq data. For microarray data see supplementary figure S3.

TDGs missed by DELocal

The DELocal algorithm appears to efficiently identify the DE genes (having high precision) as well as to filter out non TDGs (having high specificity). Still, DELocal missed 60 differentially expressed TDGs which are identified by all the other methods in RNAseq dataset. DELocal is built on the hypothesis that every true DE gene should have neighbours and none of them should be differentially expressed. Consequently, DELocal may fail to identify those DE genes whose neighbours are also differentially expressed. For instance, 2 out of 5 neighbours of *Bmp7* were DE genes which could be the reason of failure of DELocal to detect *Bmp7* (Figure 6). Additionally, presence of paralogous genes in the neighbourhood may contradict with our hypothesis, like *Dlx1* and *Dlx2* in chromosome 11 (Figure 6), *Dlx5* and *Dlx6* in chromosomes 6, and *Dlx1* and *Dlx2* in chromosome 2, for which it is known that the pairs of these genes are co-regulated and can compensate for the deletion of one another (29-32). Additionally, the other neighbouring paralogs in our set of TDGs are *Cyp26c1* with *Cyp26a1*, and *Cdh1* with *Cdh3* (33). Therefore, at least some of the genes missed by DELocal should be possible to detect by incorporating information about gene paralogues into the algorithm. These results also raise the question whether our tooth example is representative about organ regulation in general.

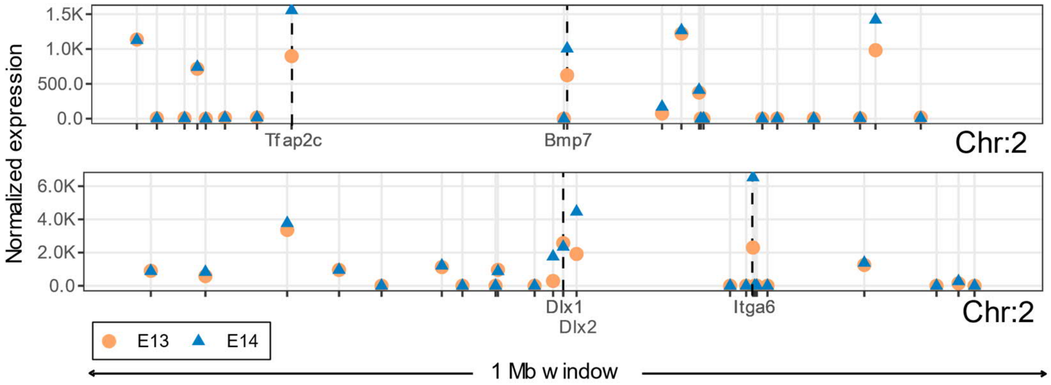

Figure 6. Differentially expressed genes in the same neighbourhood interfere with the detection of differentially expressed genes by DELocal. DELocal failed to identify *Bmp7* due to differential expression patterns of its neighbouring genes (top)**.** *Dlx1* and *Dlx2 are p*aralogous TDGs which are in the same neighbourhood (bottom). Only tooth developmental genes are labelled here.

Potential of DELocal beyond tooth genes

The DELocal method presented here depends on the existence of neighbouring genes. For applications beyond our example, it is interesting to examine more closely the composition of 1 Mb neighbourhoods in the mouse genome. Majority of the protein-coding genes from Affymatrix MoEX chip have more than one neighbour within 1 Mb, and only 7 genes in ensemble mouse annotation lack neighbours altogether within 1 Mb window (Figure 1A). If genes, which are involved in the same developmental process, had a tendency to cluster in the same genomic location, then the expression pattern of that neighbourhood would become dynamic and the outcome of DELocal would deteriorate. Thus, DELocal's performance will depend on the distribution of relevant genes across chromosomal locations. Although it is difficult to conduct an experiment for all the possible functional groups of genes, the gene ontology (GO) terms provides a rough approximation of the adjacency of genes involved in the same developmental process. Here we focused in the genes belonging to mouse GO slims (a list of selected terms) of "biological process" (34,35). For every gene belonging to these terms, its 1 Mb neighbourhood was investigated to tabulate genes belonging to the same term. The tabulations show that most genes belonging to a certain GO term are sparsely distributed in the chromosomes (Figure 7) which is in accordance with previous studies (36,37). There are only few genes from broader, or high level, GO terms that are densely located (more than 3 genes from the same GO term within 1 Mb neighbourhood). However, these few GO terms represent very broad descriptions of biological functions. With more precise GO terms (Figure 7, bottom row), genes tend to have no neighbours belonging to the same GO term. For generality, we examined these patterns in human genome and they remained largely the same (Supplementary Figure S4).

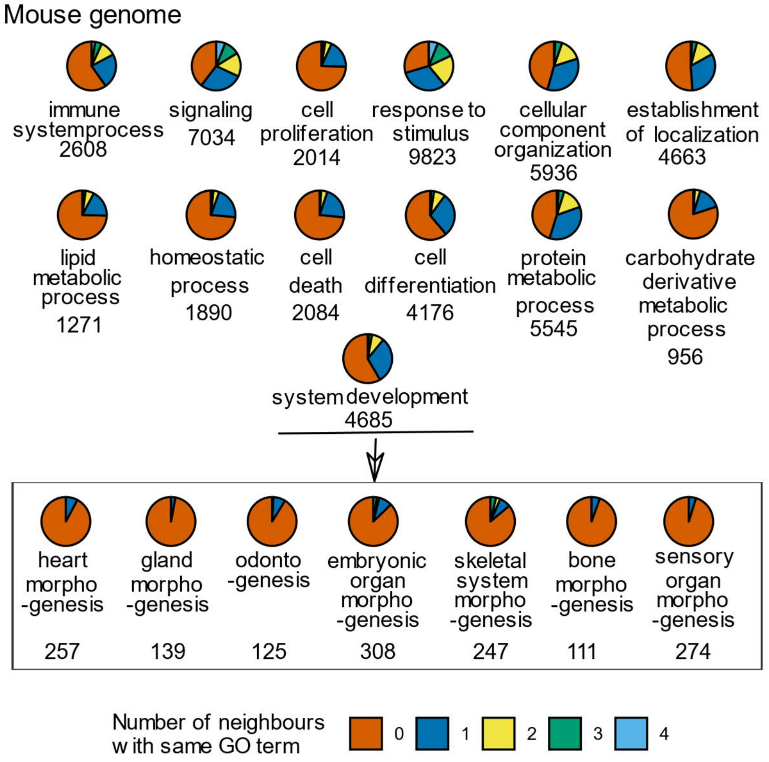

Figure 7. The more specific the GO term is the fewer neighbours its genes have from the same GO term within 1 Mb. Each pie represents the genes of one GO term under the root GO term ‘biological processes’. The top three rows represent GO slim terms. The GO terms in the box are children of the GO term ‘system development’. The color-coding indicates the number of neighbours that a gene has from the same GO term. The number of genes is indicated under the GO term. Analysis was done for mouse genome. See Supplementary Table S2 for GO IDs.

Simulations

The scarcity of neighbouring genes in the precise GO term categories is perhaps not surprising considering the limited number of genes in each category. To test if the patterns are a simple function of group size and to evaluate to what extent neighbouring genes might potentially interfere with the DELocal analysis, we performed simulations where genes were assigned randomly to artificial GO terms containing different numbers of genes. For every group size, 10,000 simulations were performed and for each simulation density of genes from the same artificial GO term in 1Mb neighbourhood was calculated. A plot showing the percentage of genes lacking neighbours shows the expected decrease in percentage as the number of genes increases in the artificial GO terms (Figure 8). The empirical patterns largely follow the randomizations, although some real GO terms with small number of genes (300 and less) show slightly lower percentages, indicating that there is tendency for higher spatial clustering of genes belonging to the same GO terms than in the simulations. This may partly be due to

GO terms having paralogous genes which are sometimes located near each other in the genome. Nevertheless, up to GO term categories containing 1000 genes, 80 % of genes have no neighbours belonging to the same GO terms. This pattern indicates that DELocal or equivalent methods could be broadly applicable.

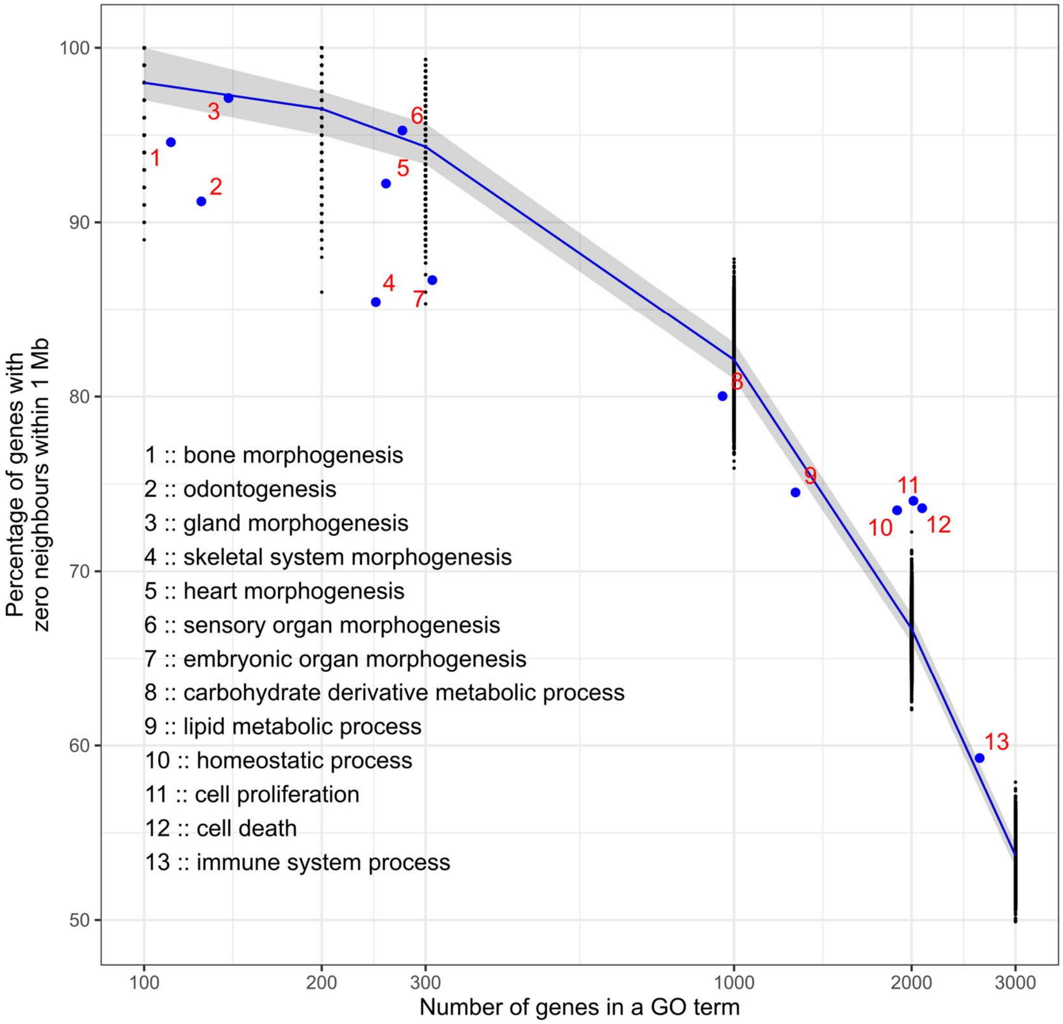

Figure 8. Real GO terms with lower gene numbers are slightly more clustered in the genome than artificial GO terms with the same gene numbers. Artificial GO terms with different numbers of genes were made with randomly selected genes and their distribution across the genome was measured. For each group size, 10,000 simulations were executed and for each simulation the percentage of genes with zero within-1Mb-neighbours with the same artificial GO term were counted (black dots). Real GO terms are marked with blue dots. See Supplementary Table S2 for GO IDs. The blue line is median, and the shadow shows the observations between 1st and 3rd quartile.

Conclusions

Here we developed DELocal algorithm based on linear models that have been successfully implemented and used in Limma, DeSeq2 and many other methods to identify DE genes (11,13). With linear models, gene expression can be modelled in two or more biological conditions and thereafter DE genes of different contrasts of interest can be found. Linear models are advantageous compared to other methods in that they can model complex experimental conditions with multiple factors. DELocal provides some additional benefits by taking into account the gene neighbourhood of genes of interest. In this work we used DELocal to determine the genes that are differentially expressed between bud and cap stage in the developing mouse tooth. Earlier work has produced an extensive list of genes which are active in tooth development. This information allowed us to optimize and evaluate the performance of DELocal. DELocal provided high specificity and accuracy in detecting TDGs, a satisfactory result. Considering that the in vivo bud and cap stage differences in gene expression are relatively subtle, the high specificity and accuracy of DELocal is promising. Obviously, the optimization requires a list of genes of interest. However, DELocal can also be run without any prior knowledge of genes that are active in a particular developmental process. Figure 3 shows that in relation to the number of neighbours, the performance of DELocal is very stable and even only one nearest neighbour could be sufficient to build the models. The implication of this is that DELocal could be used with only a single neighbour contrast in the absence of any reference/training gene set.

A key requirement for this approach is the assumption that genes participating in the same function tend to be far from each other (Figure 1). That this is indeed the case is suggested by an analysis using GO terms (Figure 7, 8) that shows the overall majority of neighbouring genes to belong to different categories. Thus, neighbourhood information on differentially expressed genes should be applicable to majority of developmental systems and processes.

There are a few groups of genes which are clustered in certain chromosomal regions and their expression is regulated in a concerted manner, for example the Hox genes or immunoglobulin genes. Also house-keeping genes have been shown to cluster in the genome (37,38). As the operation of DELocal depends on the non-differential expression of the neighbours, the potential differential expression of clustered genes cannot be analysed by DELocal.

Overall, DELocal is a novel way to reveal differentially expressed genes with respect to their genomic neighbourhood. Future studies may benefit from, and further characterize the significance of gene-gene neighbour relationships.

FUNDING

Financial support was provided by the Academy of Finland, Jane and Aatos Erkko Foundation, John Templeton Foundation, and Sigrid Jusélius Foundation.

Conflict of interest statement. None declared.

DATA AVAILABILITY

The DELocal algorithm is freely available here https://github.com/dasroy/DELocal as an R package. The microarray and RNAseq data are deposited in the NCBI Gene Expression Omnibus with accession codes GSE141907 and GSE142199.

ACKNOWLEDGEMENTS

We thank Pasi Rastas for discussions. We thank P. Auvinen, L. Paulin and P. Laamanen at DNA Sequencing and Genomics Laboratory for bulk RNA sequencing.

REFERENCES

1. Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, **3**, 318-356.
2. Zabidi, M.A. and Stark, A. (2016) Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors. *Trends in genetics : TIG*, **32**, 801-814.
3. Dong, X., Li, C., Chen, Y., Ding, G. and Li, Y. (2010) Human transcriptional interactome of chromatin contribute to gene co-expression. *BMC genomics*, **11**, 704.
4. Thévenin, A., Ein-Dor, L., Ozery-Flato, M. and Shamir, R. (2014) Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic Acids Research*, **42**, 9854-9861.
5. Vogel, J.H., von Heydebreck, A., Purmann, A. and Sperling, S.J.B.B. (2005) Chromosomal clustering of a human transcriptome reveals regulatory background. **6**, 230.
6. Lemay, D.G., Martin, W.F., Hinrichs, A.S., Rijnkels, M., German, J.B., Korf, I. and Pollard, K.S.J.B.B. (2012) G-NEST: a gene neighborhood scoring tool to identify co-conserved, co-expressed genes. **13**, 253.
7. Hallikas, O., Das Roy, R., Christensen, M., Renvoisé, E., Sulic, A.-M. and Jernvall, J. (2020) System-level analyses of keystone genes required for mammalian tooth development. *J Exp Zool Part B Mol Dev Evol [in press]*.
8. Del Carratore, F., Jankevics, A., Eisinga, R., Heskes, T., Hong, F. and Breitling, R. (2017) RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. *Bioinformatics (Oxford, England)*, **33**, 2774-2775.
9. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116-5121.
10. Ilmjärv, S., Hundahl, C.A., Reimets, R., Niitsoo, M., Kolde, R., Vilo, J., Vasar, E. and Luuk, H. (2014) Estimating differential expression from multiple indicators. *Nucleic Acids Research*, **42**, e72-e72.
11. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47-e47.

12. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, **26**, 139-140.
13. Love, M.I., Huber, W. and Anders, S.J.G.B. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **15**, 550.
14. Bengtsson, H., Irizarry, R., Carvalho, B. and Speed, T.P. (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics (Oxford, England)*, **24**, 759-767.
15. Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, **33**, e175.
16. Andrews, S. (2010), Vol. 2019.
17. Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M. and Gu, J.J.B.B. (2017) AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. **18**, 80.
18. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, **30**, 2114-2120.
19. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, **29**, 15-21.
20. Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, **31**, 166-169.
21. Risueno, A., Roson-Burgo, B., Dolnik, A., Hernandez-Rivas, J.M., Bullinger, L. and De Las Rivas, J. (2014) A robust estimation of exon expression to identify alternative spliced genes applied to human tissues and cancer samples. *BMC genomics*, **15**, 879.
22. Nieminen, P., Pekkanen, M., Aberg, T. and Thesleff, I. (1998) A graphical WWW-database on gene expression in tooth. *Eur J Oral Sci*, **106 Suppl 1**, 7-11.
23. Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, **21**, 3940-3941.
24. Durinck, S., Spellman, P.T., Birney, E. and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, **4**, 1184-1191.
25. Lane, N. and Martin, W. (2010) The energetics of genome complexity. *Nature*, **467**, 929-934.
26. Jernvall, J. and Thesleff, I. (2012) Tooth shape formation and tooth renewal: evolving with the same signals. *Development*, **139**, 3487-3497.
27. Maija, K., Pekka, N., Carin, S., Thomas, B. and Irma, T. (1996-2007). University of Helsinki.
28. Li, C.Y., Hu, J., Lu, H., Lan, J., Du, W., Galicia, N. and Klein, O.D. (2016) alphaE-catenin inhibits YAP/TAZ activity to regulate signalling centre formation during tooth development. *Nature communications*, **7**, 12133.
29. Thomas, B.L., Tucker, A.S., Qui, M., Ferguson, C.A., Hardcastle, Z., Rubenstein, J.L. and Sharpe, P.T. (1997) Role of Dlx-1 and Dlx-2 genes in patterning of the murine dentition. *Development*, **124**, 4811-4818.
30. Qiu, M., Bulfone, A., Ghattas, I., Meneses, J.J., Christensen, L., Sharpe, P.T., Presley, R., Pedersen, R.A. and Rubenstein, J.L. (1997) Role of the Dlx homeobox genes in proximodistal patterning of the branchial arches: mutations of Dlx-1, Dlx-2, and Dlx-1 and -2 alter morphogenesis of proximal skeletal and soft tissue structures derived from the first and second arches. *Dev Biol*, **185**, 165-184.

31. Beverdam, A., Merlo, G.R., Paleari, L., Mantero, S., Genova, F., Barbieri, O., Janvier, P. and Levi, G. (2002) Jaw transformation with gain of symmetry after Dlx5/Dlx6 inactivation: mirror of the past? *Genesis*, **34**, 221-227.
32. Debiais-Thibaud, M., Metcalfe, C.J., Pollack, J., Germon, I., Ekker, M., Depew, M., Laurenti, P., Borday-Birraux, V. and Casane, D. (2013) Heterogeneous conservation of Dlx paralog co-expression in jawed vertebrates. *PloS one*, **8**, e68182.
33. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M. *et al.* (2015) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, **44**, D286-D293.
34. The Gene Ontology, C. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*, **47**, D330-D338.
35. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**, 25-29.
36. de Laat, W. and Grosveld, F. (2003) Spatial organization of gene expression: the active chromatin hub. *Chromosome Research*, **11**, 447-459.
37. Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature genetics*, **31**, 180-183.
38. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A. *et al.* (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289-1292.