BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 4 2013, pages 468–475
doi:10.1093/bioinformatics/bts720

Genetics and population analysis

Advance Access publication December 24, 2012

Disease-gene discovery by integration of 3D gene expression and transcription factor binding affinities

Rosario M. Piro[1,2,*], Ivan Molineris[3,4], Ferdinando Di Cunto[3,4], Roland Eils[1,2] and Rainer König[1,2]

[1]Department of Theoretical Bioinformatics, German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ), [2]Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, Bioquant, University of Heidelberg, Im 69120 Heidelberg, Germany, [3]Molecular Biotechnology Center and [4]Department of Molecular Biotechnologies and Health Sciences, University of Torino, 0126 Torino, Italy

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: The computational evaluation of candidate genes for hereditary disorders is a non-trivial task. Several excellent methods for disease-gene prediction have been developed in the past 2 decades, exploiting widely differing data sources to infer disease-relevant functional relationships between candidate genes and disorders. We have shown recently that spatially mapped, i.e. 3D, gene expression data from the mouse brain can be successfully used to prioritize candidate genes for human Mendelian disorders of the central nervous system.

Results: We improved our previous work 2-fold: (i) we demonstrate that condition-independent transcription factor binding affinities of the candidate genes' promoters are relevant for disease-gene prediction and can be integrated with our previous approach to significantly enhance its predictive power; and (ii) we define a novel similarity measure—termed *Relative Intensity Overlap*—for both 3D gene expression patterns and binding affinity profiles that better exploits their disease-relevant information content. Finally, we present novel disease-gene predictions for eight loci associated with different syndromes of unknown molecular basis that are characterized by mental retardation.

Contact: r.piro@dkfz.de or rmpiro@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 31, 2012; revised on December 14, 2012; accepted on December 19, 2012

1 INTRODUCTION

Although experimental methodologies for the identification of disease-causing mutations have significantly improved—in particular by the recently introduced next-generation sequencing techniques—computational approaches to an *in silico* evaluation of candidate genes remain an important aid for the identification of genes involved in human hereditary disorders (Kann, 2010; Piro *et al.*, 2011; Piro and Di Cunto, 2012). A large variety of computational methods have been developed for this purpose, exploiting various data sources ranging from MEDLINE abstracts, functional gene annotation, protein–protein interactions and high-throughput gene expression data to intrinsic gene or protein properties (such as coding sequence length, number of introns, conservation and so forth). For more information on available disease-gene prediction tools and methods, we refer to some recent reviews (Kann, 2010; Piro and Di Cunto, 2012).

To a large extent, cellular homeostasis depends on coordinated gene expression, both in space and time. Gene expression determines *where* and *when* the molecular function of a gene product is exerted. For this reason, gene expression patterns have been successfully exploited by several computational disease-gene prediction methods (Kann, 2010; Piro and Di Cunto, 2012).

In our previous work (Piro *et al.*, 2010), we have shown that the high-resolution 3D gene expression patterns provided by the Allen Institute's Mouse Brain Atlas (MBA) (Lein *et al.*, 2007) can be used to successfully prioritize not only candidate genes for mouse phenotypes but also for human Mendelian disorders of the central nervous system (CNS). Briefly, we prioritized the candidate genes from a mapped 'orphan' locus of a disease phenotype with unknown molecular basis by comparing their 3D gene expression patterns to those of a set of 'reference genes', known to be involved in similar disease phenotypes.

The condition and sample dependency of gene expression profiles constitutes a desired feature for the prediction of functional relationships between genes. A complementary view, and hence additional clues to gene functions, may instead be provided by condition-independent regulatory information, obtained from promoter analysis (Werner, 2003). Classically, position frequency matrices (PFMs) or their associated position weight matrices (PWMs) that describe transcription factor (TF) binding sites are used to scan promoters for short, often degenerate, regulatory elements. In many cases, genome-wide PWM scans require stringent criteria to limit the false-discovery rate and to yield an acceptable specificity. This can lead to reduced sensitivity and likely misses low-affinity binding sites (Hannenhalli, 2008), which have been shown to be abundant and functional *in vivo* (Tanay, 2006).

An alternative approach builds on the notion of an overall likelihood that a promoter is bound by a given TF, without requiring the identification of one or more well-defined regulatory elements (Foat *et al.*, 2006; Molineris *et al.*, 2011; Ward and Bussemaker, 2008), thus better reflecting the thermodynamic nature of TF binding (Segal *et al.*, 2008). For this purpose, a *total binding affinity* (TBA) for the TF's PFM can be calculated over the entire promoter, taking into account the collective

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

contribution of (potential) high- and low-affinity binding sites, without strictly drawing a line between binding and non-binding sites. In a recent study, we showed TBA to be predictive of TF-binding events, as revealed by ChIP/chip and ChIP/seq experiments (Molineris *et al.*, 2011).

Here, we introduce the notion of TF binding *co-affinity* of a pair of genes that we compute from the genes' TBA profiles composed of their binding affinities for a core set of TFs. We show that these co-affinities can decipher the disease relevance of candidate genes for human hereditary disorders and further improve the predictive power by developing a new similarity measure termed *Relative Intensity Overlap* (*RIO*). We combine this approach with the prioritization based on spatially mapped, i.e. 3D, gene expression, applying it to several disorders of unknown molecular basis characterized by mental retardation (MR).

2 METHODS

2.1 Disease phenotypes and gene–disease associations

Information on human Mendelian disease phenotypes was obtained from the Online Mendelian Inheritance in Man (OMIM) database (Amberger *et al.*, 2009) on August 15, 2011, limited to those that contain the term 'central nervous system' in their clinical synopsis section and have at least one mapped disease locus. Note that this selection criterion does not exclude other organs and tissues from being affected, i.e. symptoms need not be limited to the CNS. A total of 867 OMIM phenotype entries with known molecular basis (OMIM symbol: #) were downloaded. Associated disease-related genes were determined by merging information from the OMIM Morbidmap and Entrez Gene (Sayers *et al.*, 2012) (mim2gene), yielding 948 CNS-related gene–phenotype (*g–p*) pairs. Additionally, we obtained all phenotype entries for various syndromes with an unknown molecular basis (OMIM symbol: %) that contain the term 'mental retardation' (MR) in their title field and have been mapped to a disease locus containing a set of candidate genes.

As our aim was to predict the most likely candidates for mapped 'orphan' loci from OMIM phenotype entries with so far unknown molecular basis, in which no genes are known to be involved, we used disease-associated genes from similar phenotype entries as reference genes for the prediction procedure described later in the text. For this purpose, we measured the pairwise similarity of OMIM phenotypes using MimMiner, essentially as described by van Driel *et al.* (2006). MimMiner scores are normalized and range from 0 (unrelated) to 1 (highly related or identical). Instead of the minimum score of 0.4 proposed by van Driel *et al.* to define phenotype similarity, we applied a more stringent threshold of 0.5 to focus on more relevant reference genes.

For the *ex novo* candidate gene prioritizations that we present here, we restricted our analysis to eight representative orphan disease loci for both X-linked and autosomal MR that had at least two reference genes from similar phenotypes and at least 20 candidate genes for which both gene expression and affinity data were available (see later in the text).

2.2 3D gene expression data

3D gene expression profiles for the MBA (Lein *et al.*, 2007) were downloaded on February 2, 2011 using the application programming interface provided by the Allen Institute's website at http://mouse.brain-map.org/. Only sagittal image series with antisense probes for genes with defined Entrez gene IDs were considered. In case of multiple image series per gene, only the most recent series was used. The downloaded expression patterns provide expression levels for the entire brain, smoothed over evenly spaced voxels (cubes) with a side length of 200 μm.

As we applied the mouse expression data for prioritizing human disease genes, we mapped expression profiles from the mouse brain to human Entrez gene IDs using unambiguous mappings from NCBI's HomoloGene (build 65) (Sayers *et al.*, 2012), yielding 14 590 human Entrez gene IDs with an associated expression pattern from the mouse brain. From these, we excluded 68 genes that had no corresponding TBA profiles (see later in the text).

3D expression patterns from the MBA consist only of non-negative values ($\geq$0) that we normalized as follows: for each profile, we first sorted all positive expression levels (>0) in increasing order and then set the 95th percentile to a value of 0.95 (to avoid taking outliers into account for normalization). All other expression values were scaled linearly by the same factor. Expression levels that exceeded 1 after scaling were interpreted as potential outliers and limited to equal 1.

2.3 Brain region-specific gene expression

To evaluate the performance of different correlation measures in clustering spatial expression profiles, we used sets of genes that exhibit region-specific expression in the mouse brain. Of the top 100 genes for each of 12 brain regions, defined by Lein *et al.* (2007), we retained only those that could be unambiguously mapped to human genes (see earlier in the text), obtaining between 77 and 92 (on average 84) genes per region.

We then used a rank-based procedure to construct ranked co-expression groups (RCGs; Miozzi *et al.*, 2008) for all 14 522 genes, each being composed of the gene that defines the RCG plus the k genes with the most correlated 3D gene expression profiles (where correlation is defined by one of the tested similarity measures). Varying k from 1 to 100, we scanned the RCGs for gene pairs from the 12 brain region-specific gene sets.

2.4 Total binding affinity and co-affinity

As described in our recent study on the evolution of promoter–TF affinity, for each human transcription start site, we consider the region spanning 1500 bp upstream to 500 bp downstream as promoter, but the approach is robust with respect to the choice of promoter size (Molineris *et al.*, 2011).

Following Foat *et al.* (2006), we define the TBA a_{rw} of a promoter r for a transcription factor, represented by its position frequency matrix (PFM) w, as:

$$a_{rw} = log \sum_{i=1}^{L-l} max\left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j})}{P(b, r_{i+j})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j})}{P(b, r_{i+j})}\right) \quad (1)$$

where L is the length of the promoter, l is the length of the PFM, r_i is the nucleotide at position i of the promoter and r'_i is the nucleotide at the same position on the opposite strand.

The probability $P(w_j, r_{i+j})$ to observe a given nucleotide r_{i+j} at promoter position $i+j$ also at position j of the PFM w is computed as the PFM's frequency count of the nucleotide at w_j, divided by the total frequency count for all four nucleotides at w_j. In the case of a frequency count of zero, we add a pseudocount of 1. $P(w_j, r_{i+j})$ is corrected for the background probability $P(b, r_{i+j})$ of r_{i+j}, computed as the nucleotide frequency for the whole intergenic part of the human genome.

For genes with multiple promoters r (i.e. multiple transcription start sites), we set the gene's TBA a_{gw} for a given PFM w (i.e. a given transcription factor or transcription factor family) to the maximum value obtained for its individual promoters r_k, i.e.

$$a_{gw} = \max(a_{r_1 w}, \ldots, a_{r_{n(g)} w}) \quad (2)$$

where $n(g)$ is the number of promoters r of gene g. This is biologically reasonable because a single promoter of the gene is in principle sufficient to drive its expression (although the promoter may allow to produce only some of the possible isoforms of the gene).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

As affinities from different PFMs are in most cases not directly comparable [because of widely differing lengths l and nucleotide frequencies $P(w_j)$],we finally apply a z-transformation to the TBA scores a_{gw}, such that each gene g's TBA is represented by the direction and number of standard deviations it differs from the mean TBA $\bar{a}_w$ for the given PFM w.

We computed TBAs of 130 non-redundant vertebrate PFMs deposited in the Jaspar Core database (Bryne *et al.*, 2008) for 37 231 human promoters defined by transcription start sites from RefSeq (Pruitt *et al.*, 2007). We could map the RefSeq IDs to 22 120 Entrez gene IDs (using only unambiguous mappings). Consequently, each gene g can be described by a TBA profile $\vec{a}_g = (a_{gw_1}, a_{gw_2}, \ldots, a_{gw_{130}})$ composed of its (z-transformed) affinities for the transcription factors represented by the PFMs [Equations (1) and (2)]. However, we limit the evaluation presented here to those 14 522 genes, i.e. TBA profiles, for which MBA profiles are also available (see earlier in the text).

Similar to the co-expression between two genes that can be estimated from their gene expression profiles, we define the *co-affinity* of two genes as the similarity of their TBA profiles (with different possible measures of similarity). To this notion of co-affinity, we can apply the same candidate gene prioritization and leave-one-out cross validation (LOOCV) procedures defined for gene expression (see later in the text).

2.5 Relative intensity overlap

We define the RIO of two 3D gene expression profiles a and b (that can be thought of as 3D images) as follows:

$$\mathrm{RIO}(a,b) = \frac{\sum_{xyz}(I^a_{xyz} \cdot I^b_{xyz})}{\sum_{xyz}\left(\max(|I^a_{xyz}|, |I^b_{xyz}|)\right)^2} \tag{3}$$

where I^a_{xyz} and I^b_{xyz} are intensities, i.e. expression levels, at voxel xyz for genes a and b, respectively, and sums $\left(\sum_{xyz}\right)$ are calculated over all voxels.

Interpretation: The RIO measures the overlap of two (normalized) 3D gene expression profiles by multiplying them with each other and summing the contributions of the single voxels to an overall score. This score is then normalized through division by the maximum possible overlap that would be obtained if both profiles had for each voxel the higher expression level of the two. This way, a maximum contribution to the score can be obtained for $I^a_{xyz} = I^b_{xyz} = 1$, whereas voxels with low expression values ($I^a_{xyz} = I^b_{xyz} \approx 0$) will neither give significant contributions nor are explicitly penalized. Voxels with maximally discordant expression (e.g. $I^a_{xyz} = 1$ and $I^b_{xyz} = 0$), instead, will yield no contribution but maximum penalization [maximum summand for the denominator in Equation (3)].

Analogously, we define the RIO of two TBA profiles m and n as in Equation (3), using as intensities the (z-transformed) TBAs a_{mw} and a_{nw} and summing over the PFMs w instead of summing over voxels.

Note that RIOs can range from 0 to 1 if the intensities are non-negative quantities (as is the case for the MBA expression data) or from -1 to 1 if the intensities can include negative values (as is the case for the z-transformed TBAs) because in the latter case, discordant intensities can yield negative contributions (if, e.g. $a_{mw} > 0$ and $a_{nw} < 0$).

2.6 Candidate gene prioritization

For both MBA and TBA profiles, we use the candidate gene prioritization method, outlined in Figure 1, that we defined in our previous study for 3D gene expression profiles (Piro *et al.*, 2010).

Briefly, for each reference gene $r \in R_p$, selected for a given OMIM disease phenotype p (via MimMiner; see earlier in the text), all other genes are ranked according to the similarity of their MBA or TBA profiles, respectively, to obtain one genome-wide, ranked co-expression or co-affinity list for each r (columns in Fig. 1). For this purpose, profile similarity can be defined in various ways, for example, by the Pearson correlation coefficient [PCC; used in Piro *et al.* (2010) for the MBA data] or the RIO that we propose here (see earlier in the text).

The given set of candidate genes C_p (e.g. candidates from an orphan disease locus) has now to be prioritized, i.e. ranked, according to their likelihood of being involved in p. We apply the following procedure: the rank/position $k(c,r)$ of each candidate gene $c \in C_p$ within each of the ranked profile similarity lists (columns) of the reference genes r is determined, and a relative rank $k(c,r)/k_{max}$ computed, where k_{max} is the total number of genes in the ranked lists. Subsequently, each candidate gene is assigned an overall score s_c, determined as the product of its relative ranks within the reference genes' co-expression or co-affinity lists,

$$s_c = \prod_{r \in R_p} \frac{k(c,r)}{k_{max}} \tag{4}$$

and candidates are sorted, i.e. prioritized, according to their increasing overall score, thus giving precedence to lower scores (as lower scores indicate better rankings in the single profile similarity lists and, therefore, a higher probability of being functionally related to the reference genes). For more details, see Piro *et al.* (2010).

2.7 Leave-one-out cross validation

To demonstrate the validity of our approach—including the new similarity measure and the prediction based on affinity profiles—we performed large-scale LOOCVs for all known gene–disease phenotype associations (g–p pairs) regarding CNS-related Mendelian disorders from OMIM (see earlier in the text). For each g–p pair, we removed all gene–disease associations of p and constructed an 'artificial locus' containing the disease-related gene g itself plus the N closest genes on both sides of the chromosome (for $N = 50$, $N = 100$, $N = 200$ and $N = 400$), hence, simulating an orphan locus of at most $2N + 1$ genes (or less for g close to a chromosome terminal) linked to an OMIM phenotype of unknown molecular basis. Then, we prioritized the candidates from the artificial loci and verified the absolute rank ($\mathcal{R}_g$) and relative rank $\mathcal{R}^{rel}_g$ (=rank $\mathcal{R}_g$ divided by the number of candidates) of the true disease-related gene g within the prioritized candidate list.

The analysis was limited to gene–phenotype pairs whose corresponding artificial loci contain at least 50 'effective' candidate genes for which both MBA and TBA profiles are available—one of which was required to be g itself—as only these can be evaluated and thus prioritized. For a lower number of effective candidate genes, an undesired bias would tend to automatically place the true phenotype-causing gene g in high ranks $\mathcal{R}_g$. See Table 1 for the number of gene–phenotype pairs that could be evaluated.

2.8 Integration of candidate rankings

To obtain overall rankings for a given set of candidate genes, we first compute distinct candidate rankings for the MBA-based and the TBA-based approach, respectively, as illustrated in Figure 1. For each candidate gene c, we determine its (predicted) ranks $\mathcal{R}^M_c$ and $\mathcal{R}^T_c$ from the two ranked candidate lists (M = MBA, T = TBA). Using an adaptation of the generalized noisy-OR gate defined by Díez (1993), we compute the conditional probability that c 's true rank $\mathcal{R}^*_c$ is 1 (i.e. c is the true disease gene) as

$$P(\mathcal{R}^*_c = 1 | \mathcal{R}^M_c, \mathcal{R}^T_c) = P(\mathcal{R}^*_c = 1 | \mathcal{R}^M_c) \cdot P(\mathcal{R}^*_c = 1 | \mathcal{R}^T_c) \tag{5}$$

i.e. the probability that c is the true disease gene conditionally on having been ranked at position $\mathcal{R}^M_c$ for the MBA, multiplied by the probability that it is the true disease gene given that it has been ranked at position $\mathcal{R}^T_c$ for the TBA. We determine the probability $P(\mathcal{R}^*_c = 1 | \mathcal{R}^D_c)$ for each single dataset $D \in (M, T)$ empirically from the leave-one-out performance over all evaluated gene–phenotype pairs (g, p) as

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

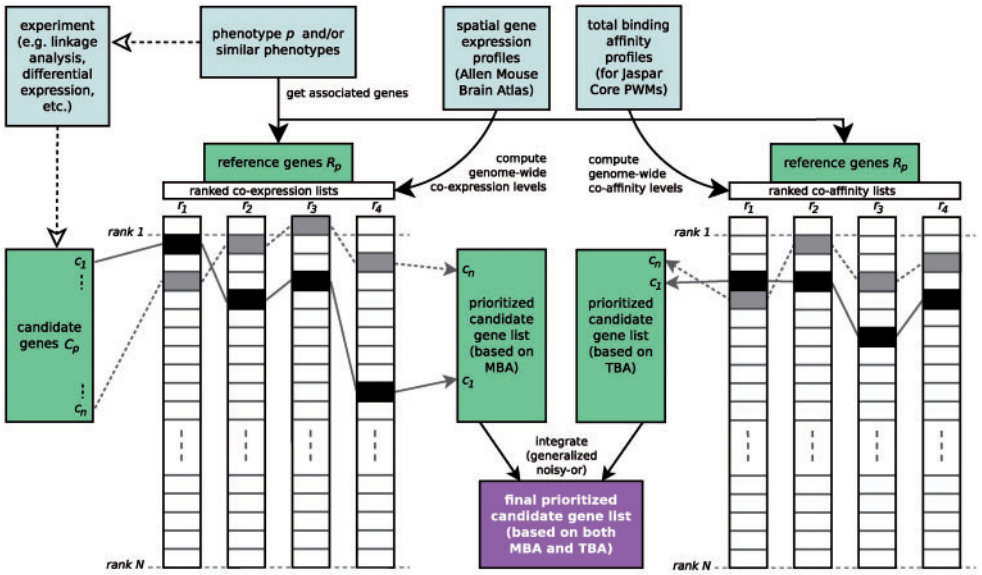

Fig. 1. Schematic representation of the disease-gene prioritization procedure, exemplified with two of the hypothetical candidate genes c_i and four reference genes r_i associated to a phenotype p. The procedure is first applied individually to MBA profiles and TBA profiles, and the resulting prioritized candidate lists are integrated into an overall ranking. The locus containing the candidate genes C_p, the disease phenotype p and the expression and affinity profiles are considered as given [adapted from Piro *et al.* (2010)]

Table 1. Results of the LOOCV

Sim.	N	C_p	g–p pairs	Ranked first			Ranked 1st–3rd			Ranked 1st–10th			Ranked≤10%		
				Obs.	E.	P-value	Obs.	E.	P-value	Obs.	E.	P-value	Obs.	E.	P-value
(a) —TBA; RIO versus PCC															
PCC	50	73.4	756	17	10	3.23e-02	52	31	2.20e-04	138	103	2.01e-04	103	76	8.17e-04
PCC	100	136.3	805	9	6	1.42e-01	33	18	6.08e-04	87	59	2.15e-04	112	81	2.41e-04
PCC	200	253.3	808	3	3	6.19e-01	17	10	1.80e-02	60	32	3.34e-06	111	81	4.19e-04
PCC	400	439.3	808	3	2	2.80e-01	11	6	2.52e-02	41	18	2.69e-06	118	81	2.26e-05
RIO	50	73.4	756	25[a]	10	5.62e-05	58[a]	31	4.48e-06	175[a]	103	7.76e-13	125[a]	76	1.54e-08
RIO	100	136.3	805	19[a]	6	1.15e-05	37[a]	18	3.05e-05	116[a]	59	3.34e-12	136[a]	81	1.06e-09
RIO	200	253.3	808	13[a]	3	2.70e-05	21[a]	10	8.46e-04	69[a]	32	3.23e-09	146[a]	81	2.01e-12
RIO	400	439.3	808	10[a]	2	2.16e-05	16[a]	6	1.90e-04	46[a]	18	2.69e-08	151[a]	81	5.81e-14
(b)—**MBA**; RIO versus PCC															
PCC	50	73.4	756	22	10	8.53e-04	59	31	2.19e-06	157	103	3.73e-08	119	76	4.82e-07
PCC	100	136.3	805	16	6	3.81e-04	39	18	5.81e-06	97	59	1.25e-06	123	81	1.63e-06
PCC	200	253.3	808	9	3	5.40e-03	26	10	6.81e-06	62	32	8.01e-07	126	81	4.40e-07
PCC	400	439.3	808	8	2	6.19e-04	17	6	5.93e-05	38	18	3.20e-05	131	81	2.90e-08
RIO	50	73.4	756	29[a]	10	8.62e-07	58	31	4.48e-06	156	103	6.30e-08	119	76	4.82e-07
RIO	100	136.3	805	18[a]	6	3.91e-05	39	18	5.81e-06	102[a]	59	6.23e-08	121	81	4.41e-06
RIO	200	253.3	808	13[a]	3	2.70e-05	26	10	6.81e-06	58	32	1.30e-05	125	81	7.42e-07
RIO	400	439.3	808	9[a]	2	1.22e-04	13	6	4.38e-03	41[a]	18	2.69e-06	129	81	8.93e-08
(c)—Integrated approach (TBA + MBA); RIO															
RIO	50	73.4	756	**38**	10	9.78e-12	**97**	31	3.33e-23	**230**	103	2.20e-33	**174**	76	6.31e-26
RIO	100	136.3	805	16	6	3.81e-04	**58**	18	7.85e-15	**166**	59	6.18e-34	**206**	81	5.63e-37
RIO	200	253.3	808	**20**	3	1.70e-10	**50**	10	8.09e-21	**140**	32	8.48e-49	**262**	81	2.03e-68
RIO	400	439.3	808	**16**	2	1.16e-10	**43**	6	1.89e-24	**124**	18	1.70e-62	**315**	81	1.81e-105

LOOCV results obtained with the PCC and the RIO as similarity measure (Sim.) for (a) TBA-based predictions and (b) MBA-based predictions. (c) Results obtained for the integrated approach (TBA + MBA). N represents the size of the artificial loci having a maximum of $2N+1$ genes. The average numbers of effective candidates C_p with both TBA and MBA profiles and the numbers of evaluated g–p pairs are shown. The observed (Obs.) and expected (E.) numbers of g–p pairs, for which the true phenotype-causing gene g ranks first, among the top three, among the top 10 and within the best 10% of the prioritized list, are reported along with the corresponding P-values (one-tailed Fisher's exact test).

Grey background highlights the results obtained with our previous method (upper box) and the new integrated approach (lower box). Bold face font highlights the improvements for this comparison (improvement for all results but one; none worsened).

[a]Results for which the RIO outperforms the PCC.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

$$P(\mathcal{R}_c^* = 1|\mathcal{R}_c^D) = \frac{\left|\{(g,p)|\mathcal{R}_g^D = \mathcal{R}_c^D\}\right| + 1}{\left|\{(g,p)\}\right|} \quad (6)$$

corresponding to the fraction of evaluated gene–phenotype pairs for which the true disease gene g ranked $\mathcal{R}_c^D$, like the candidate gene c. To prevent probabilities from becoming nil, we add a pseudocount of one.

Finally, we determine an integrated candidate gene list by prioritizing the candidates c according to their decreasing probability [Equation (5)] of being the true disease gene. For disambiguation, candidate genes with equal probabilities $P(\mathcal{R}_c^* = 1|\mathcal{R}_c^M, \mathcal{R}_c^T)$ are further prioritized according to their ranks $\mathcal{R}_c^M$ and $\mathcal{R}_c^T$ for the individual datasets (TBA or MBA), giving precedence to candidates with better single rankings.

3 RESULTS

3.1 Promoter–TF binding affinities are predictive

We performed LOOCVs over all known CNS-related gene–disease associations to verify whether TBA profiles can actually aid in prioritizing candidate genes from known disease loci. At the same time, we asked whether the frequently used PCC or our new similarity measure, the RIO, would better suit this task. For this purpose, we constructed artificial candidate loci of various sizes around each known disease gene and compared the candidates' TBA profiles with those of related reference genes, using either PCC or RIO as similarity measure.

As demonstrated in Table 1a, TBA profiles are indeed predictive and can be used to successfully prioritize candidate genes. Furthermore, our new similarity measure RIO outperforms the PCC for all locus sizes and for all evaluated categories of disease-gene rankings (first, top three, top 10 and best 10%). The number of disease genes correctly being ranked first, for example, is increased by 47–333% depending on the locus size. Likewise, RIO outperforms both the Spearman rank correlation (SRC) and the cosine-similarity (COS), as shown in Supplementary Table S1a. Thus, for comparing binding affinity profiles, we chose to use our new similarity measure.

3.2 RIO for predictions based on 3D gene expression

The measurement of profile similarity is of major interest also when applying 3D gene expression patterns to evaluate which candidate genes are more likely to be functionally associated with a given disorder. We asked whether RIO outperforms PCC—used both in our previous work (Piro *et al.*, 2010) and by the Allen Institute (Lein *et al.*, 2007)—in clustering spatial expression profiles and predicting likely candidates for CNS-related disorders. In addition, we compared the results with those obtained for SRC and COS.

To evaluate the performance in clustering spatial expression profiles, we constructed RCGs (Miozzi *et al.*, 2008) of varying sizes $k+1$ ($k = 1$ to 100) for all genes and all four profile similarity measures, and we counted the number of pairs of genes within these RCGs known to be specifically expressed in the same brain region (see Section 2). Figure 2 shows that for the MBA, RIO can consistently better recover region-specific relationships between gene expression patterns. For RCGs consisting only of each gene plus it's genome-wide most correlated gene ($k = 1$), for example, RIO identifies 233 gene pairs specifically expressed in the same brain region, whereas PCC, COS and SRC identify only 199, 180 and 153, respectively.

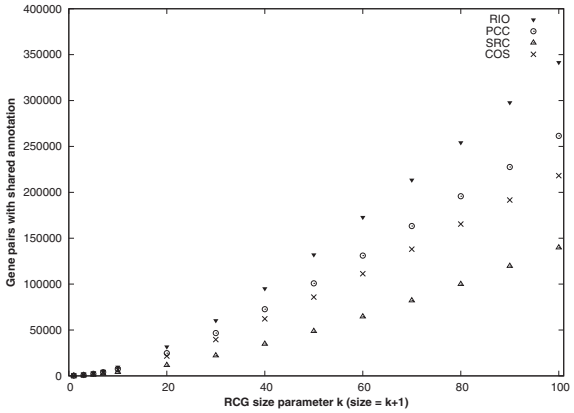

Fig. 2. Performance of RIO, PCC, SRC and COS in clustering spatial expression profiles. The number of observed gene pairs with shared annotation regarding brain region-specific gene expression is plotted over the RCGs' size parameter k (see Section 2)

For evaluating RIO's impact on predicting likely candidates for CNS-related disorders, we performed additional LOOCVs. As shown in Table 1b, RIO performs comparably with PCC with respect to true disease-associated genes that are ranked among the first three, first 10 or the best 10% of the candidate genes from artificial loci. Furthermore, when considering disease-genes that are ranked first among the set of candidates—the ones that would preferably be evaluated by geneticists on the search for causes of hereditary disease—RIO clearly outperforms PCC (increase of 12–44%). In this regard, it also compares favourably with both SRC and COS, as shown in Supplementary Table S1b. Therefore, in the following sections, we used our new similarity measure also for this type of gene expression data.

However, the performance improvement of the RIO depends on a proper normalization of the 3D gene expression data. For the LOOCV results for raw MBA profiles along with a discussion, see the Supplementary Information (Supplementary Table S2 and accompanying text).

3.3 Complementarity of MBA and TBA

As the MBA gene expression data and the TBAs regard two different species (mouse and human) and consider two largely different aspects of gene expression (dynamic, tissue- and location-specific expression levels versus a static, condition-independent predisposition for binding events), we reasoned that a joint application of both could be beneficial for the purpose of prioritizing candidate genes for hereditary disorders.

For the RIO and the smallest artificial loci ($N = 50$), for example, the disease-associated gene g ranked first or second for 45 g–p pairs with the MBA-based and for 42 g–p pairs with the TBA-based approach, respectively, but there was no overlap at all between these correct predictions (see Supplementary Table S3). Overall, median Spearman correlation coefficients of the leave-one-out rankings obtained for MBA and TBA were close to zero, ranging from 0.031 (for $N = 50$) to 0.044 (for $N = 400$). This indicates that successful predictions of the MBA-based approach are usually not obtained through the TBA-based approach and vice versa, providing a strong rationale for an integration of the two approaches.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Therefore, we integrated the results obtained for the two individual prioritizations using an adaptation of the generalized noisy-OR gate (see Section 2). We found that their complementarity considerably increases the predictive power with respect to our previous work (Piro *et al.*, 2010). This becomes clear when comparing the results for MBA-based predictions with PCC (grey box in Table 1b)—reflecting the same method that we used for our previous study—with those for the new combined approach (Table 1c). Indeed, for the different locus sizes and evaluation categories, the number of recovered disease genes increased by up to 226% (for $N = 400$ and ranking among the top 10).

3.4 Comparison to other prioritization methods

Börnigen *et al.* (2012) have recently evaluated eight commonly used candidate gene prioritization methods and obtained realistic performance estimates for *ex novo* predictions by applying the tools to novel gene–disease associations within a few days of their publication. They obtained areas under the receiver operating characteristic curves (AUCs) of 0.56–0.86.

In comparison, for our unbiased LOOCV over the largest artificial loci ($N = 400$), containing the most candidates, we obtained an AUC of 0.81 (see Supplementary Fig. S1). Only two of the tools tested by Börnigen *et al.* (2012), GeneDistiller (Seelow *et al.*, 2008) and Endeavour (Aerts *et al.*, 2006) obtained slightly higher AUCs (0.86 and 0.83, respectively). Hence, the performance of our approach is generally in line with the most efficient approaches commonly being used.

3.5 Novel predictions for mental retardation

Given the positive outcome of the large-scale LOOCV that simulates the application of our method to OMIM phenotypes of unknown molecular basis (OMIM symbol: %), we applied the procedure to the eight partly overlapping orphan loci listed in Table 2 that are involved in various syndromes (both autosomal and X-linked) of which MR is an important clinical feature.

4 DISCUSSION

Regulatory information from PWM scans has previously been used as an *additional* information source in a number of candidate gene prioritization methods (Kann, 2010; Piro and Di Cunto, 2012). However, because of the stringent criteria that are often necessary to decide whether a specific DNA segment is a binding site, leading to reduced sensitivity and a likely disregard of possibly functional low-affinity binding sites (Hannenhalli, 2008), it is unlikely that information from PWM scans *alone* would be suitable for disease-gene discovery. Likewise, the benefit of more comprehensive TBA profiles is not immediately clear and has previously not been explored on a large scale.

Although gene function can often be inferred from primary coding sequences, deciphering functional properties *exclusively* from non-coding sequences is much more difficult (Carroll, 2005). For one, co-affinity is static and condition-independent; hence, it does not directly translate into the co-expression of, say, two genes in brain tissues. In contrast to gene expression data, the concept of binding affinity does not implicitly integrate information on different types of regulatory control (such as the methylation state of promoters, nucleosome positioning and the post-transcriptional downregulation by microRNAs). Instead, co-affinities, being calculated from promoter sequences only, are limited to the direct binding of TFs to promoters.

Second, even a successful binding event must not necessarily alter transcription rates if, for example, required co-factors are lacking in a specific tissue. Indeed, Gao *et al.* (2004) suggest that as much as 42% of TF binding may be non-functional in yeast.

Finally, binding affinity profiles constitute a description of a *cis*-regulatory sequence, in this case, the promoter, involved in the regulatory control of a gene (Molineris *et al.*, 2011). But given the incompleteness of the TBA profiles we considered (composed of a core set of 130 PWMs; see Section 2) or, more general, the incompleteness of our knowledge regarding functional TF binding, a successful application of this kind of information to disease-gene prediction was not necessarily guaranteed (for a more detailed discussion, see Supplementary Text S1).

Nonetheless, in this study, we could show that regulatory information alone, in the form of TBA-based co-affinities, can indeed prioritize candidate genes for hereditary disorders. Further improvements of the predictive power were obtained by the application of a new similarity measure, RIO, and by integration with our previous approach based on 3D gene expression patterns (Piro *et al.*, 2010). We exploited the new integrated approach to identify promising candidates for several MR syndromes (Table 2).

Strikingly, among the best candidates for most of the orphan loci, we found genes that have been reported to be implicated in one or more other MR-related syndromes (Table 2 and Supplementary Table S5). Notably, several of these MR genes obtained high ranks, although they were not included in our original set of known gene–disease associations (and, hence, in our set of possible reference genes). For example, *synaptophysin* (*SYP*) is involved in MR, X-linked 96 (OMIM ID #300802) (Tarpey *et al.*, 2009), whose phenotype entry is brief and does not contain the term 'central nervous system' in its clinical synopsis. Nonetheless, we found the gene among the top 3% candidates for the largest orphan locus, cubitus valgus with MR and unusual facies (OMIM %300471), underlining that our method is indeed capable to identify promising candidate genes that are not yet known to be linked to MR. Likewise, *ZMYM3*, our top candidate for the Prieto X-linked MR syndrome (OMIM %309610), is not associated to any CNS-related phenotype entry in OMIM, although a chromosomal translocation (X;13) involving *ZMYM3* is associated with X-linked MR (van der Maarel *et al.*, 1996).

Additionally to known MR genes, some of our top ranking candidates have already been suggested as possibly being involved in MR syndromes or have been implicated in, or proposed for, other neurological diseases (including neurodegenerative disorders like Alzheimer's disease; Table 2 and Supplementary Table S5).

For space reasons, here, we discuss only the top candidates for alopecia/MR syndrome 1 (APMR1; %203650), originally described by Baraitser *et al.* (1983).

Genetic syndromes characterized by alopecia (hair loss) and severe MR are rare disorders of largely unknown molecular basis. In 2006, by studying a large consanguineous family from Pakistan, John *et al.* (2006) mapped the APMR1 locus to the 5.4 Mb region on chromosome 3 located between markers D3S1232 and D3S2436. Of the candidates at this locus, they screened for

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 2. Novel predictions for MR syndromes with unknown molecular basis

OMIM disease phenotype and locus	Best candidates			
Alopecia-MR syndrome 1 (%203650; 3q26.3–q27.3) 46 candidates	1. *CLCN2*[a]	4. *DVL3*	7. ***MCCC1***[b]	10. *EIF4A2*
	2. *ETV5*	5. *ABCF3*	8. *TBCCD1*	
	3. *DCUN1D1*[a]	6. *PSMD2*	9. *ST6GAL1*	
MR, X-linked, syndromic 11 (%300238; Xq26–q27) 59 candidates	1. *ZNF280C*	4. ***GRIA3***[b,c]	7. *AIFM1*[a]	10. *FHL1*
	2. ***ARHGEF6***[b]	5. *SLC25A14*[c]	8. *CD40LG*	
	3. ***SMARCA1***	6. *HTATSF1*	9. *ZIC3*[c]	
MR, X-linked, with short stature (%300360; Xq24) 32 candidates	1. ***UPF3B***[b]	4. *THOC2*	7. ***UBE2A***[b]	10. *SH2D1A*
	2. ***CUL4B***[b]	5. *NKRF*	8. *IL13RA1*	
	3. *NKAP*	6. *LONRF3*	9. *AKAP14*	
Cubitus valgus with MR and unusual facies (%300471, chromosome X) 504 candidates	1. ***SYN1***[a]	6. *DCAF12L2*	11. *ABCD1*[a]	16. *PIM2*
	2. ***WDR13***	7. *TCEAL3*	12. ***MED12***[a,b]	17. *SYTL5*
	3. *APOO*	8. ***PTCHD1***[a,b]	13. *ARMCX6*	18. ***ATP6AP2***[b]
	4. *HCFC1*	9. *SH3KBP1*[c]	14. *RNF113A*	19. ***TAZ***
	5. *DNASE1L1*	10. *SLC25A14*[c]	15. ***SYP***[b]	20. *MPP1*
Prieto X-linked MR syndrome (%309610; Xp11–q21) 109 candidates	1. ***ZMYM3***[b]	6. ***EFNB1***	11. ***SMC1A***[b]	16. *VSIG4*
	2. ***KDM5C***[b]	7. *PDZD11*	12. ***OPHN1***[b]	17. *EDA2R*
	3. *CITED1*	8. *RGAG4*	13. *OTUD6A*	18. *FOXR2*
	4. ***SLC16A2***[b]	9. *ITM2A*	14. ***FGD1***[b]	19. ***FAM123B***
	5. *ITGB1BP2*	10. ***DLG3***[b]	15. *GJB1*[a]	20. *NLGN3*[a]
Pachygyria with MR, seizures and arachnoid cysts (%600176; 11p15) 187 candidates	1. *SMPD1*[a]	6. *ART1*	11. *DEAF1*	16. ***HRAS***[b]
	2. *AP2A2*	7. *FAM160A2*	12. *MYOD1*	17. *TNNI2*
	3. *PHLDA2*	8. *RASSF7*	13. *PHRF1*	18. *TNNT3*
	4. *MICAL2*	9. *RNH1*	14. *STIM1*	19. *OR51I1*
	5. *SBF2*	10. *RRM1*	15. *WEE1*	20. *DKK3*
Cerebellar ataxia, MR and dysequilibrium syndrome 2 (%610185; 17p13) 120 candidates	1. *CAMKK1*	6. *TSR1*	11. ***PFN1***	16. *CTDNEP1*
	2. *ARRB2*[c]	7. *C17orf81*	12. *SLC43A2*	17. *SLC2A4*
	3. *ABR*	8. *SGSM2*	13. *KIF1C*[c]	18. *NEURL4*
	4. *ENO3*	9. *DHX33*	14. *GGT6*	19. *VPS53*
	5. *SMYD4*	10. *PITPNA*	15. *RABEP1*	20. *RPAIN*
Kahrizi syndrome (%612713; 4p12-q12) 42 candidates	1. *KIT*	4. *SPINK2*	7. *SLAIN2*	10. *GSX2*
	2. *CEP135*	5. *ZAR1*	8. *RASL11B*	
	3. *FIP1L1*	6. *CNGA1*	9. *OCIAD2*	

Best candidates for orphan loci from OMIM associated with autosomal and X-linked MR phenotypes. Gene map loci are as reported by OMIM, but whenever possible more accurate chromosomal locations were taken (Supplementary Information and Supplementary Table S4). The number of evaluated candidate genes is reported along with the 10 best ranking candidate genes. For larger loci (with >100 candidates), the top 20 candidates are listed. Known or potential MR genes are highlighted in bold. For associated disorders and references, see Supplementary Information (Supplementary Table S5). For complete prioritized candidate lists, see Supplementary Information (Supplementary Tables S6–S13).

[a]Genes known to be involved in other neurological disorders.

[b]Genes known to be involved in MR-related syndromes.

[c]Genes potentially involved in other neurological disorders.

mutations in coding exons only the transcription factor *ETV5*, but no abnormalities were found (John *et al.*, 2006). Interestingly, we found *ETV5* as the second best candidate. However, although it is in principle possible that mutations affecting its regulatory regions may be responsible for the phenotype, this seems unlikely on the basis of the available functional information. Indeed, *ETV5* loss of function has been shown to affect the maintenance of spermatogonial stem cells (Chen *et al.*, 2005) and kidney development (Lu *et al.*, 2009), whereas *ETV5* overexpression has been linked to ovarian cancer (Llauradó *et al.*, 2012).

The only top 10 candidate that has been linked to an MR-related disorder is *MCCC1*, mutations of which are known to be responsible for 3-methylcrotonyl-CoA carboxylase 1 deficiency (Baumgartner *et al.*, 2001; Gallardo *et al.*, 2001), which is characterized by multiple clinical features, one of which can be MR (Murayama *et al.*, 1997; Steen *et al.*, 1999). Alopecia, however, is generally not a feature of the disease, although a possible association has been reported in a single case (Leonard *et al.*, 1981).

However, as a particular strength of the predictive method that we present here is an extrapolation of functional relationships from reference genes to candidate genes that—relying on gene expression and sequence information alone—do not require any known functional annotation, less obvious candidates could nonetheless be associated with the disorder.

The first candidate, *CLCN2*, encodes a voltage-gated chloride channel, previously implicated in idiopathic generalized epilepsies (Haug *et al.*, 2003). *DCUN1D1*, ranking third, is another interesting candidate because a polymorphism has been identified as a risk factor in frontotemporal lobar degeneration (Villa *et al.*, 2009). Also *DVL3*, ranking forth, is of particular interest because the protein it encodes interacts with Shank, which is involved in several neuronal disorders including MR (Saupe *et al.*, 2011).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

A final promising candidate may be *ADIPOQ* (ranking 18th; Supplementary Table S6), although adiponectin, the protein it encodes, is thought to be expressed specifically in adipose tissue (Hu *et al.*, 1996). Nonetheless, recent research indicates that the secreted protein influences the proliferation of adult hippocampal neural stem/progenitor cells that express adiponectin receptors 1 and 2 (AdipoR1 and AdipoR2) (Zhang *et al.*, 2011), and adiponectin deficiency could also negatively impact brain microcirculation (Vachharajani *et al.*, 2012). Moreover, a low-protein level and a mutation of the corresponding gene were found in a patient affected by Werner syndrome (Hashimoto *et al.*, 2007), which is characterized, among other features, by prominent hair loss.

We think that these examples show well how the predictions obtained by our approach could support geneticists to dissect the molecular basis of many ill-defined human disorders.

ACKNOWLEDGEMENT

The authors thank Paolo Provero at the University of Torino, Italy, for his precious suggestions that helped to improve this work.

Funding: Nationales Genom-Forschungs-Netz (NGFN+), project ENGINE (01GS0898); Helmholtz Alliance on Systems Biology (SB Cancer, D.141100/07.997); CancerSys-Verbundprojekt: MYC-NET (0316076C); Regione Piemonte; FIRB-Italbionet program of the Italian Ministry of Education, University and Research; DKFZ Postdoctoral Fellowship (to R.M.P.).

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.

Amberger,J. *et al.* (2009) McKusick's online mendelian inheritance in man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.

Baraitser,M. *et al.* (1983) A new alopecia/mental retardation syndrome. *J. Med. Genet.*, **20**, 64–65.

Baumgartner,M.R. *et al.* (2001) The molecular basis of human 3-methylcrotonyl-CoA carboxylase deficiency. *J. Clin. Invest.*, **107**, 495–504.

Börnigen,D. *et al.* (2012) An unbiased evaluation of gene prioritization tools. *Bioinformatics*, **28**, 3081–3088.

Bryne,J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.

Carroll,S.B. (2005) Evolution at two levels: on genes and form. *PLoS Biol.*, **3**, e245.

Chen,C. *et al.* (2005) ERM is required for transcriptional control of the spermatogonial stem cell niche. *Nature*, **436**, 1030–1034.

Diez,F.J. (1993) Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In: *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, pp. 99–105.

Foat,B.C. *et al.* (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–e149.

Gallardo,M.E. *et al.* (2001) The molecular basis of 3-methylcrotonylglycinuria, a disorder of leucine catabolism. *Am. J. Hum. Genet.*, **68**, 334–346.

Gao,F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.

Hashimoto,N. *et al.* (2007) A patient with Werner syndrome and adiponectin gene mutation. *Diabetes Res. Clin. Pract.*, **75**, 27–29.

Haug,K. *et al.* (2003) Mutations in CLCN2 encoding a voltage-gated chloride channel are associated with idiopathic generalized epilepsies. *Nat. Genet.*, **33**, 527–532.

Hu,E. *et al.* (1996) AdipoQ is a novel adipose-specific gene dysregulated in obesity. *J. Biol. Chem.*, **271**, 10697–10703.

John,P. *et al.* (2006) Localization of a novel locus for alopecia with mental retardation syndrome to chromosome 3q26.33-q27.3. *Hum. Genet.*, **118**, 665–667.

Kann,M.G. (2010) Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief. Bioinform.*, **11**, 96–110.

Lein,E.S. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.

Leonard,J.V. *et al.* (1981) Inherited disorders of 3-methylcrotonyl CoA carboxylation. *Arch. Dis. Child.*, **56**, 53–59.

Llauradó,M. *et al.* (2012) ETV5 transcription factor is overexpressed in ovarian cancer and regulates cell adhesion in ovarian cancer cells. *Int. J. Cancer*, **130**, 1532–1543.

Lu,B.C. *et al.* (2009) Etv4 and Etv5 are required downstream of GDNF and Ret for kidney branching morphogenesis. *Nat. Genet.*, **41**, 1295–1302.

Miozzi,L. *et al.* (2008) Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data. *PLoS ONE*, **3**, e2439.

Molineris,I. *et al.* (2011) Evolution of promoter affinity for transcription factors in the human lineage. *Mol. Biol. Evol.*, **28**, 2173–2183.

Murayama,K. *et al.* (1997) Isolated 3-methylcrotonyl-CoA carboxylase deficiency in a 15-year-old girl. *Brain Dev.*, **19**, 303–305.

Piro,R.M. *et al.* (2010) Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR. *Bioinformatics*, **26**, i618–i624.

Piro,R.M. *et al.* (2011) An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. *Eur. J. Hum. Genet.*, **19**, 1173–1180.

Piro,R.M. and Di Cunto,F. (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, **279**, 678–696.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Saupe,J. *et al.* (2011) Discovery, structure-activity relationship studies, and crystal structure of nonpeptide inhibitors bound to the Shank3 PDZ domain. *ChemMedChem*, **6**, 1411–1422.

Sayers,E.W. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.

Seelow,D. *et al.* (2008) GeneDistiller–distilling candidate genes from linkage intervals. *PLoS One*, **3**, e3874.

Segal,E. *et al.* (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, **451**, 535–540.

Steen,C. *et al.* (1999) Metabolic stroke in isolated 3-methylcrotonyl-CoA carboxylase deficiency. *Eur. J. Pediatr.*, **158**, 730–733.

Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.*, **16**, 962–972.

Tarpey,P.S. *et al.* (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat. Genet.*, **41**, 535–543.

Vachharajani,V. *et al.* (2012) Adiponectin-deficiency exaggerates sepsis-induced microvascular dysfunction in the mouse brain. *Obesity (Silver Spring)*, **20**, 498–504.

van der Maarel,S.M. *et al.* (1996) Cloning and characterization of DXS6673E, a candidate gene for X-linked mental retardation in Xq13.1. *Hum. Mol. Genet.*, **5**, 887–897.

van Driel,M.A. *et al.* (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Villa,C. *et al.* (2009) DCUN1D1 is a risk factor for frontotemporal lobar degeneration. *Eur. J. Neurol.*, **16**, 870–873.

Ward,L.D. and Bussemaker,H.J. (2008) Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics*, **24**, i165–i171.

Werner,T. (2003) Promoters can contribute to the elucidation of protein function. *Trends Biotechnol.*, **21**, 9–13.

Zhang,D. *et al.* (2011) Adiponectin stimulates proliferation of adult hippocampal neural stem/progenitor cells through activation of p38 mitogen-activated protein kinase (p38MAPK)/glycogen synthase kinase 3 (GSK-3)/-catenin signaling cascade. *J. Biol. Chem.*, **286**, 44913–44920.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016