Supplementary Materials to:

Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR

Rosario M. Piro, Ivan Molineris, Ugo Ala, Paolo Provero, Ferdinando Di Cunto
Molecular Biotechnology Center and Department of Genetics, Biology and Biochemistry
University of Torino, Italy – corresponding author: rosario.piro@unito.it

A. Supplementary Methods

Translation of Ensembl gene IDs to Entrez gene IDs

Ensembl gene IDs reported for the genes on chromosome X resequenced by Tarpey et al. [1] have been mapped to Entrez gene IDs using Ensembl 55 [2].

Mouse brain microarray expression data (GNF)

For comparison with microarray expression data, we used the GNF expression atlas of Su et al. [3], taking however only the samples from mouse brain tissues (8x2 data points for 8 tissues with 2 replicates). Only probesets with unique mappings to Entrez gene IDs (original GNF annotation) were used and the expression profiles of multiple probesets for the same gene were averaged, yielding expression profiles for a total of 12,676 Entrez mouse genes.

B. Supplementary Tables

Table S1: Results of the leave-one-out tests for human phenotypes with (at least partly) "known molecular basis" (OMIM class: #) using ABA expression data, and for mouse phenotypes using the GNF brain microarray data. *N* represents the size of the artificial loci having at most 2*N*+1 genes. The average numbers of effective candidates with expression data and the numbers of evaluated *g*-*p* pairs are shown. The observed and expected numbers of *g*-*p* pairs, for which the true phenotype-causing gene *g* ranks first, among the top ten and within the best 10% of the prioritized list, is reported along with the corresponding p-values (one-tailed Fisher exact test). Significant p-values are highlighted (*<0.05; **<0.01; ***<0.001).

organism, phenotypes	*N*	candidates (average)	*g*-*p* pairs	ranked first			ranked 1st–10th			ranked ≤ 10%		
				obs.	exp.	p-value	obs.	exp.	p-value	obs.	exp.	p-value
human, molecular basis known (#)	50	73.8	797	16	11	7.95e-02	150	108	1.75e-05***	127	80	1.04e-07***
human, molecular basis known (#)	100	137.7	844	13	6	9.83e-03**	106	61	3.36e-08***	133	84	1.13e-07***
human, molecular basis known (#)	200	256.6	847	6	3	1.16e-01	63	33	1.21e-06***	140	85	2.82e-09***
mouse [microarray: GNF-brain]	50	52.9	61	2	1	3.21e-01	11	12	6.22e-01	10	6	8.50e-02
mouse [microarray: GNF-brain]	100	80.9	292	6	4	1.55e-01	48	36	2.38e-02*	41	29	1.64e-02*
mouse [microarray: GNF-brain]	200	143.8	311	4	2	1.72e-01	32	22	1.75e-02*	39	31	8.33e-02

Table S2: Prioritization (*evaluation*) of the chromosome X genes resequenced by Tarpey et al. [1] for XLMR, according to the overall score s_c (see Eq. 1 in the paper). XLMR was considered as phenotype with *unknown* molecular basis; known XLMR genes were considered as candidates and only genes known to be involved in similar phenotypes were taken as reference genes. Only the best 10% of the prioritized list is shown. Gene names, IDs (except Entrez), associations to disorders and mutation scores are as reported by Tarpey et al. [1]. Mutation scores reflect the conservation scores at missense positions [1] and are summed over the single missense mutations found for each gene.

Rank	Gene	Ensembl ID	Refseq ID	Entrez ID	Disorder	Mutation score	Score s_c
1	BRWD3	ENSG00000165288	NM_153252	254065	XLMR	2.86	7.16E-078
2	IRAK1	ENSG00000184216	NM_001569	3654	-	1.94	1.84E-071
3	SYP	ENSG00000102003	NM_003179	6855	XLMR	-	8.97E-069
4	BIRC4	ENSG00000101966	-	331	other	37.04	4.28E-068
5	MAGED1	ENSG00000179222	NM_001005332	9500	-	5.14	5.63E-068
6	MORF4L2	ENSG00000123562	NM_012286	9643	-	-	3.37E-066
7	ZNF280C	ENSG00000056277	NM_017666	55609	-	-	5.07E-065

Table S2 (continued)

8	SYN1	ENSG00000008056	NM_006950	6853	XLMR	-	1.15E-064
9	CXorf6	ENSG00000013619	-	10046	other	12.00	1.19E-064
10	ATP6AP2	ENSG00000182220	NM_005765	10159	XLMR	-	2.70E-064
11	HCFC1	ENSG00000172534	NM_005334	3054	-	27.82	1.65E-061
12	PJA1	ENSG00000181191	NM_022368	64219	-	2.44	1.82E-061
13	NGFRAP1	ENSG00000166681	NM_206917	27018	-	-	1.91E-061
14	FAM50A	ENSG00000071859	NM_004699	9130	-	11.62	5.63E-061
15	HUWE1	ENSG00000086758	NM_031407	10075	XLMR	46.75	1.62E-060
16	GRIA3	ENSG00000125675	NM_007325	2892	XLMR	13.00	1.14E-059
17	PIGA	ENSG00000165195	NM_002641	5277	other	-	3.24E-059
18	OGT	ENSG00000147162	NM_181672	8473	-	15.90	4.15E-059
19	GNL3L	ENSG00000130119	NM_019067	54552	-	22.49	1.68E-058
20	WDR40C	ENSG00000198354	NM_001013628	340578	-	0.22	3.35E-058
21	UTX	ENSG00000147050	NM_021140	7403	-	-	2.18E-057
22	RNF113A	ENSG00000125352	NM_006978	7737	-	6.49	7.01E-057
23	FLNA	ENSG00000196924	NM_001110556	2316	XLMR	14.57	1.01E-056
24	PCTK1	ENSG00000102225	NM_033018	5127	-	-	1.59E-056
25	MAGEE1	ENSG00000198934	NM_020932	57692	-	2.04	2.87E-056
26	PGRMC1	ENSG00000101856	NM_006667	10857	-	9.70	4.20E-056
27	ARHGEF9	ENSG00000131089	NM_015185	23229	XLMR	-	5.65E-055
28	ATP2B3	ENSG00000067842	NM_021949	492	-	4.49	1.48E-054
29	PGK1	ENSG00000102144	NM_000291	5230	XLMR	15.67	4.56E-054
30	DRP2	ENSG00000102385	NM_001939	1821	-	41.79	4.74E-054
31	ARMCX2	ENSG00000184867	NM_177949	9823	-	-	5.77E-054
32	TAZ	ENSG00000102125	NM_000116	6901	other	-	1.80E-053
33	OPHN1	ENSG00000079482	NM_002547	4983	XLMR	15.77	1.24E-052
34	GRIPAP1	ENSG00000068400	NM_020137	56850	-	8.17	2.38E-052
35	PLP2	ENSG00000102007	NM_002668	5355	-	-	8.21E-052
36	REPS2	ENSG00000169891	NM_001080975	9185	-	-	1.31E-051
37	RPS4X	ENSG00000198034	NM_001007	6191	-	-	1.74E-051
38	HPRT1	ENSG00000165704	NM_000194	3251	XLMR	-	2.32E-051
39	ARMCX1	ENSG00000126947	NM_016608	51309	-	-	3.56E-051
40	NLGN3	ENSG00000196338	NM_018977	54413	XLMR	-	4.67E-051
41	SAT1	ENSG00000130066	NM_002970	6303	other	-	5.22E-051
42	CSTF2	ENSG00000101811	NM_001325	1478	-	7.39	1.26E-050
43	USP9X	ENSG00000124486	NM_001039590	8239	-	-	1.31E-050
44	SLITRK2	ENSG00000185985	NM_032539	84631	-	9.62	2.16E-050
45	GPM6B	ENSG00000046653	NM_001001995	2824	-	-	1.41E-049
46	GJB1	ENSG00000169562	NM_001097642	2705	other	-	3.06E-049
47	SLC9A6	ENSG00000198689	NM_001042537	10479	XLMR	8.58	1.92E-048

Table S3: Prioritization (*prediction*) of the chromosome X genes resequenced by Tarpey et al. [1] for XLMR, according to the overall score s_c (see Eq. 1 in the paper). XLMR was considered as phenotype with *known* molecular basis; only genes not known to be involved in XLMR were considered as candidates and known XLMR gene were taken as reference genes; the concept of phenotype similarity was not applied. Only the best 10% of the prioritized list is shown. Gene names, IDs (except Entrez), associations to disorders and mutation scores are as reported by Tarpey et al. [1]. Mutation scores reflect the conservation scores at missense positions [1] and are summed over the single missense mutations found for each gene.

Rank	Gene	Ensembl ID	Refseq ID	Entrez ID	Disorder	Mutation score	Score s_c
1	MORF4L2	ENSG00000123562	NM_012286	9643	-	-	5.09E-099
2	PJA1	ENSG00000181191	NM_022368	64219	-	2.44	7.70E-097
3	ZNF280C	ENSG00000056277	NM_017666	55609	-	-	1.91E-093
4	MAGED1	ENSG00000179222	NM_001005332	9500	-	5.14	1.55E-091
5	MAGEE1	ENSG00000198934	NM_020932	57692	-	2.04	1.60E-085
6	BIRC4	ENSG00000101966	-	331	other	37.04	1.02E-084
7	GRIPAP1	ENSG00000068400	NM_020137	56850	-	8.17	3.13E-082
8	CXorf6	ENSG00000013619	-	10046	other	12.00	3.75E-081
9	GNL3L	ENSG00000130119	NM_019067	54552	-	22.49	4.07E-081
10	FAM50A	ENSG00000071859	NM_004699	9130	-	11.62	7.72E-081
11	PGRMC1	ENSG00000101856	NM_006667	10857	-	9.70	8.65E-081
12	GPM6B	ENSG00000046653	NM_001001995	2824	-	-	4.12E-079
13	IRAK1	ENSG00000184216	NM_001569	3654	-	1.94	8.19E-079
14	HCFC1	ENSG00000172534	NM_005334	3054	-	27.82	1.28E-078
15	PIGA	ENSG00000165195	NM_002641	5277	other	-	1.65E-078
16	RPS4X	ENSG00000198034	NM_001007	6191	-	-	4.97E-078
17	REPS2	ENSG00000169891	NM_001080975	9185	-	-	2.81E-077
18	ARMCX2	ENSG00000184867	NM_177949	9823	-	-	1.67E-075
19	DRP2	ENSG00000102385	NM_001939	1821	-	41.79	3.64E-074
20	MED14	ENSG00000180182	NM_004229	9282	-	9.54	1.41E-073
21	ARMCX1	ENSG00000126947	NM_016608	51309	-	-	2.94E-073
22	WDR40C	ENSG00000198354	NM_001013628	340578	-	0.22	5.30E-073
23	PCTK1	ENSG00000102225	NM_033018	5127	-	-	1.31E-072
24	OGT	ENSG00000147162	NM_181672	8473	-	15.90	1.02E-069
25	WBP5	ENSG00000185222	NM_016303	51186	-	-	2.71E-069
26	RNF113A	ENSG00000125352	NM_006978	7737	-	6.49	6.25E-069
27	BEX1	ENSG00000133169	NM_018476	55859	-	-	7.64E-069
28	PDZD11	ENSG00000120509	NM_016484	51248	-	-	1.55E-068
29	GPRASP2	ENSG00000158301	NM_138437	114928	-	4.00	1.70E-068
30	SLITRK2	ENSG00000185985	NM_032539	84631	-	9.62	3.49E-068
31	USP9X	ENSG00000124486	NM_001039590	8239	-	-	1.32E-067
32	CLCN4	ENSG00000073464	NM_001830	1183	-	21.36	1.12E-064
33	SAT1	ENSG00000130066	NM_002970	6303	other	-	1.19E-064
34	TCEAL1	ENSG00000172465	NM_001006640	9338	-	-	1.66E-064
35	TCEAL8	ENSG00000180964	NM_153333	90843	-	5.22	2.47E-064
36	TAZ	ENSG00000102125	NM_000116	6901	other	-	2.63E-064

Table S3 (continued)

37	BHLHB9	ENSG00000198908	NM_030639	80823	-	-	3.56E-064
38	CSTF2	ENSG00000101811	NM_001325	1478	-	7.39	1.40E-063
39	UTX	ENSG00000147050	NM_021140	7403	-	-	1.64E-063

Table S4: Area under ROC curve (AUC) for leave-one-out valdidations (see also Figure S1). *N* represents the size of the artificial loci having at most 2*N*+1 genes. AUCs are reported based on both the true disease gene's relative rank in the prioritized candidate list and its absolute rank in the list.

organism, phenotypes	*N*	AUC	
		based on relative rank	based on absolute rank
mouse	50	0.54889	0.55560
mouse	100	0.54659	0.55595
mouse	200	0.54774	0.56173
human, molecular basis unknown (%)	50	0.54944	0.55725
human, molecular basis unknown (%)	100	0.55541	0.56418
human, molecular basis unknown (%)	200	0.55747	0.56931
human, molecular basis known (#)	50	0.55099	0.55891
human, molecular basis known (#)	100	0.55694	0.56592
human, molecular basis known (#)	200	0.55881	0.57072

C. Supplementary Figures

Figure S1 (see next page): ROC curves for the leave-one-out validation. ROCs are shown based on both the true disease gene's relative rank in the prioritized candidate list and its absolute rank in the list. Only ROC curves for locus sizes of *N*=200 (max. 401 genes) are shown; as also suggested by the AUCs (see Table S4), the ROC curves for *N*=50 and *N*=100 are nearly identical.

D. Supplementary References

[1] Tarpey,P.S., *et al*. (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation, *Nature Genetics,* **41**(5), 535-543.

[2] Hubbard,T.J.P., *et al*. (2009) Ensembl 2009, *Nucleic Acids Research,* **37,** D690-D697.

[3] Su,A.I., *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A*. **101**(16): 6062–6067.

Figure S1 (see previous page for caption)