Supplementary Information

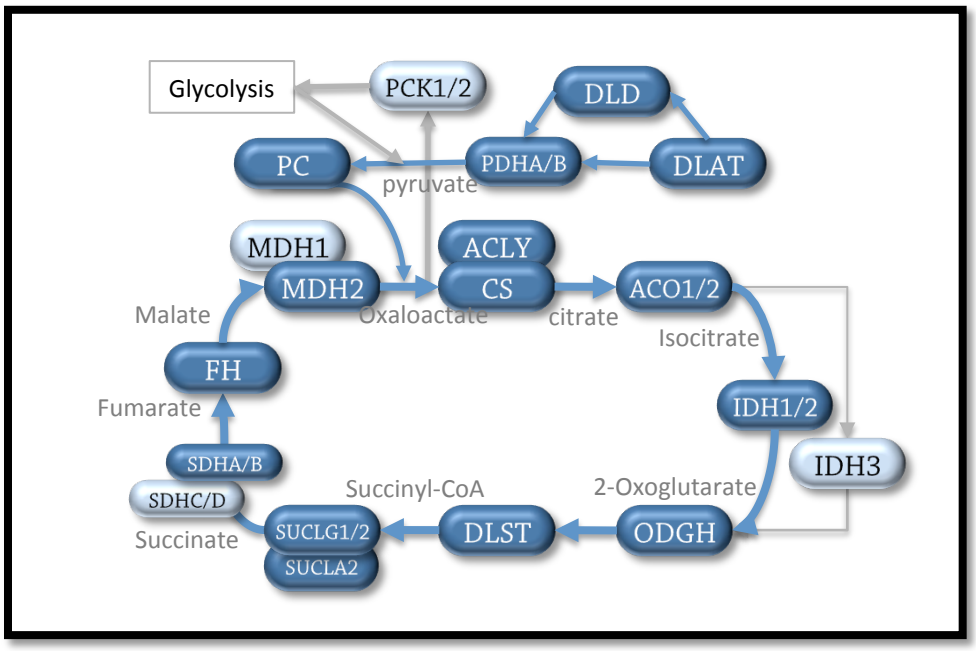
Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling

Yuval Tabach[1,2], Tamar Golan[3], Abrahan Hernández-Hernández[4], Arielle R. Messer[3], Tomoyuki Fukuda[4], Anna Kouznetsova[4], Jian-Guo Liu[4], Ingrid Lilienthal[4], Carmit Levy[3]*, Gary Ruvkun[1,2]*

[1]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114
[2]Department of Genetics, Harvard Medical School, Boston, MA 02115
[3]Department of Human Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel.
[4]Department of Cell and Molecular Biology, Karolinska Institute, Stockholm, Sweden

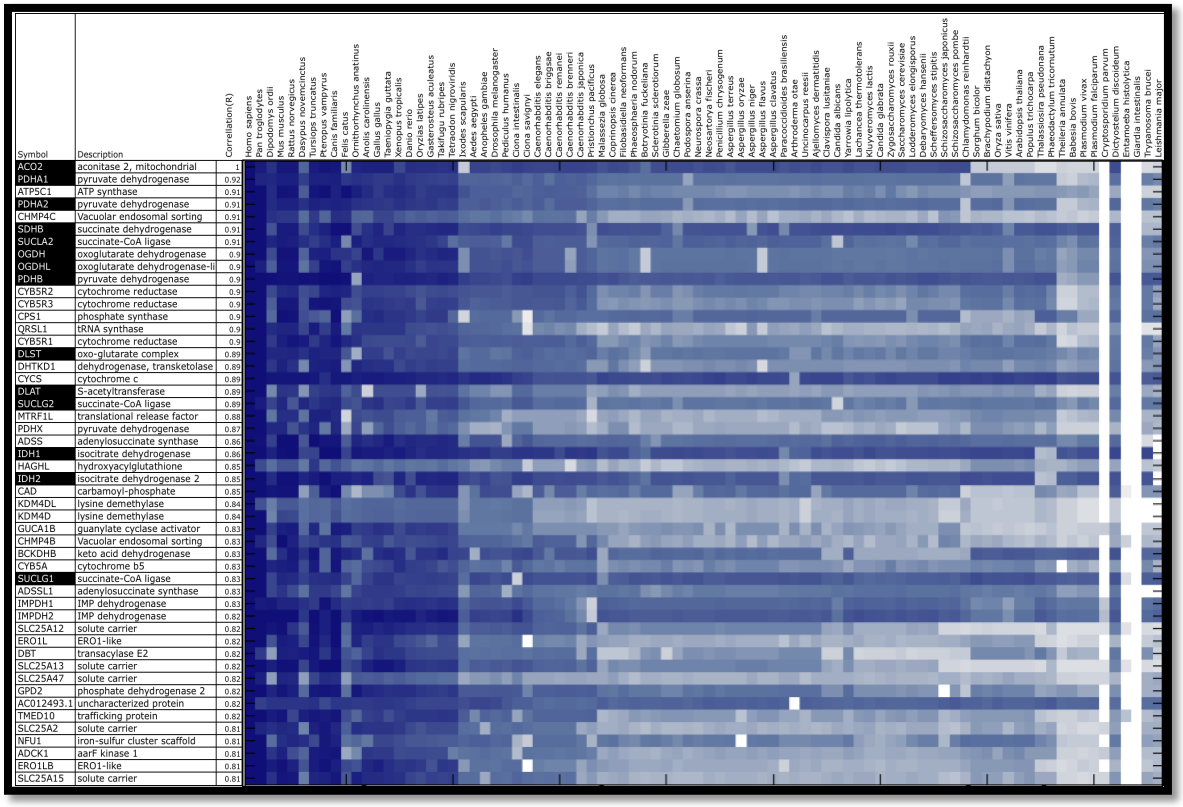
Table of content

Supplementary Figure. 1: The TCA scheme.



The blue ellipses represent genes that have homology across all or most of the organisms analyzed. The white ellipses represent genes that are conserved in particular clades of species but not others. Drawn based on KEGG (http://www.genome.jp/kegg/).

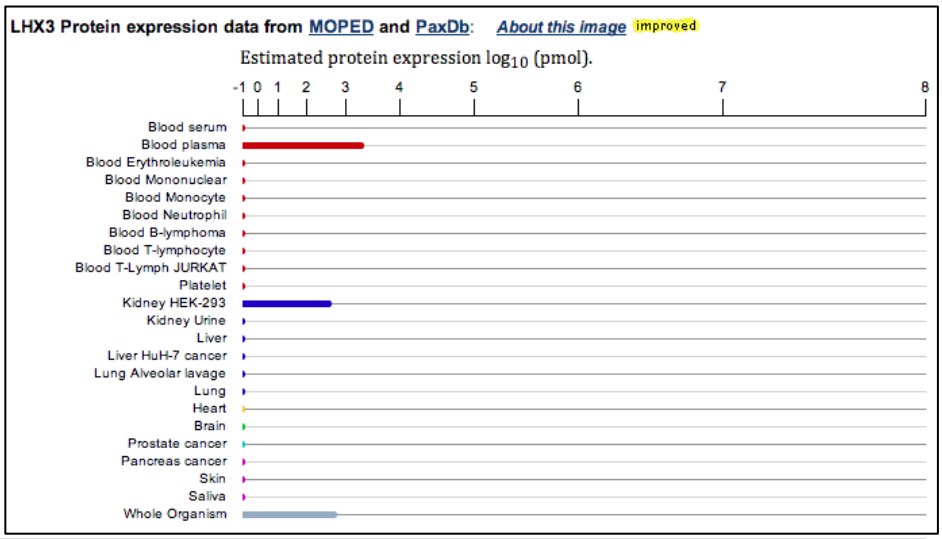
Supplementary Figure 2: The rank order of the 50 genes that are most correlated in the phylogenetic profile with the TCA gene *ACO2*.

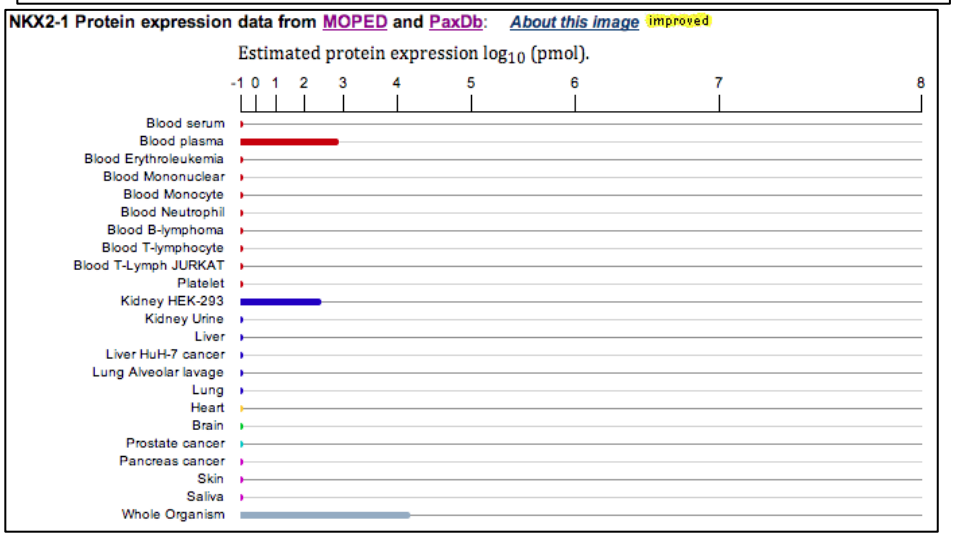


Symbol	Description	Correlation(R)
ACO2	aconitase 2, mitochondrial	1
PDHA1	pyruvate dehydrogenase	0.92
ATP5C1	ATP synthase	0.91
PDHA2	pyruvate dehydrogenase	0.91
CHMP4C	Vacuolar endosomal sorting	0.91
SDHB	succinate dehydrogenase	0.91
SUCLA2	succinate-CoA ligase	0.91
OGDH	oxoglutarate dehydrogenase	0.9
OGDHL	oxoglutarate dehydrogenase-li	0.9
PDHB	pyruvate dehydrogenase	0.9
CYB5R2	cytochrome reductase	0.9
CYB5R3	cytochrome reductase	0.9
CPS1	phosphate synthase	0.9
QRSL1	tRNA synthase	0.9
CYB5R1	cytochrome reductase	0.9
DLST	oxo-glutarate complex	0.89
DHTKD1	dehydrogenase, transketolase	0.89
CYCS	cytochrome c	0.89
DLAT	S-acetyltransferase	0.89
SUCLG2	succinate-CoA ligase	0.89
MTRF1L	translational release factor	0.88
PDHX	pyruvate dehydrogenase	0.87
ADSS	adenylosuccinate synthase	0.86
IDH1	isocitrate dehydrogenase	0.86
HAGHL	hydroxyacylglutathione	0.85
IDH2	isocitrate dehydrogenase 2	0.85
CAD	carbamoyl-phosphate	0.85
KDM4DL	lysine demethylase	0.84
KDM4D	lysine demethylase	0.84
GUCA1B	guanylate cyclase activator	0.83
CHMP4B	Vacuolar endosomal sorting	0.83
BCKDHB	keto acid dehydrogenase	0.83
CYB5A	cytochrome b5	0.83
SUCLG1	succinate-CoA ligase	0.83
ADSSL1	adenylosuccinate synthase	0.83
IMPDH1	IMP dehydrogenase	0.83
IMPDH2	IMP dehydrogenase	0.82
SLC25A12	solute carrier	0.82
ERO1L	ERO1-like	0.82
DBT	transacylase E2	0.82
SLC25A13	solute carrier	0.82
SLC25A47	solute carrier	0.82
GPD2	phosphate dehydrogenase 2	0.82
AC012493.1	uncharacterized protein	0.82
TMED10	trafficking protein	0.82
SLC25A2	solute carrier	0.81
NFU1	iron-sulfur cluster scaffold	0.81
ADCK1	aarF kinase 1	0.81
ERO1LB	ERO1-like	0.81
SLC25A15	solute carrier	0.81

The genes in the black boxes are known to be part of the TCA cycle; the white-boxed genes are predicted to interact with the TCA cycle.

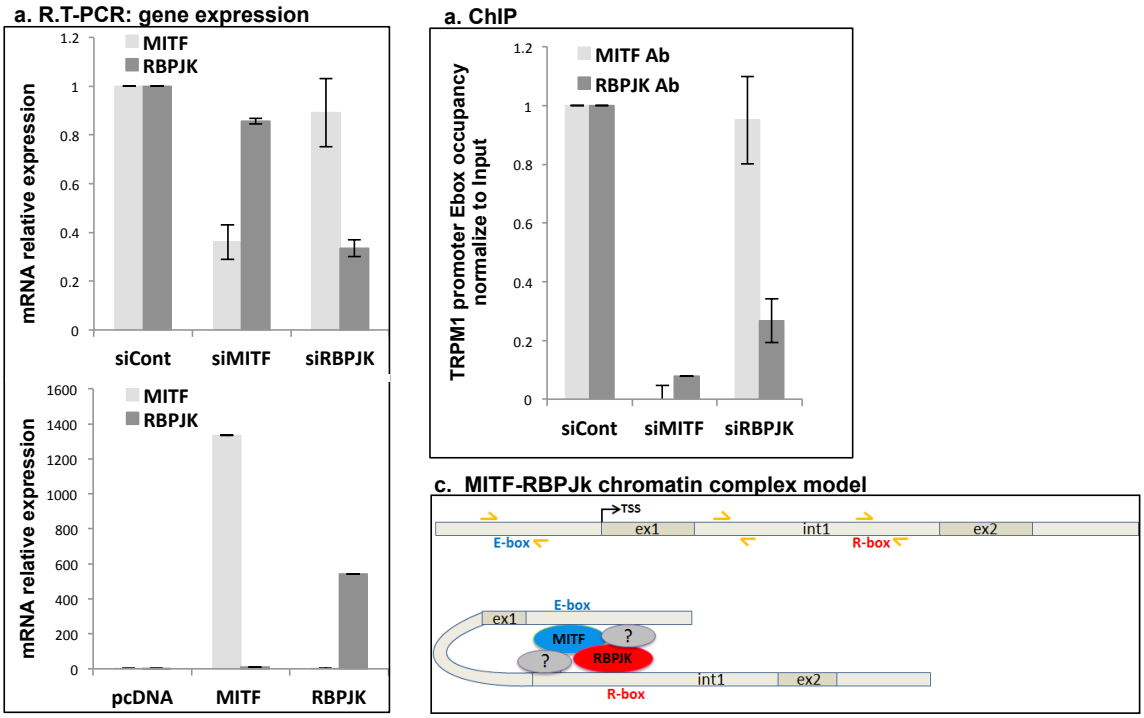
Supplementary Figure 3: LHX3 and NKX2-1 Protein expression data across different tissues.





Images were obtained from the genecards website (http://www.genecards.org/cgi-bin/carddisp.pl?gene=LHX3&search=lhx). The images demonstrate the similarity between the expression profile of LHX3 and NKX2-1 proteins across different tissues and shows high expression in liver and plasma.

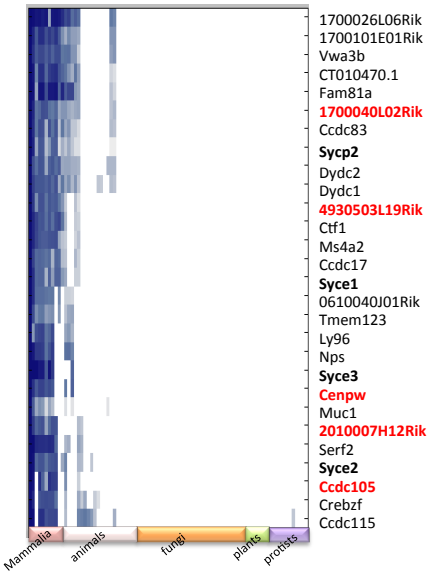
Supplementary Figure. 4: RBP-Jk is an MITF gene coregulator.



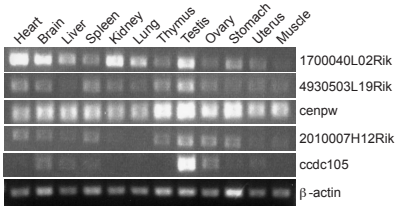
a) MITF and RBP-Jk mRNA levels were measured by RT-PCR. Primary human melanomas were transfected as indicated. MITF (dark blue bars) and RBP-Jk (light blue bars) mRNA levels were measured by qRT-PCR upon transfection, as indicated. Results are normalized to actin and are relative to siCont or empty vector respectively and represent mean ±SD of 5 replicates. **b)** MITF and RBP-Jk ChIP on *TRPM1* locus was performed upon MITF or RBP-Jk depletion. PCR primers spanning the *TRPM1* locus were employed. **c)** Diagram of RBP-Jk and MITF complex on promoter. Upper panel: *TRPM1* promoter region including DNA binding site of MITF (E-box, red). RBP-Jk (R-box), exons (ex1, ex2), intron (int1), and transcription start site (TSS) are indicated. The E-box represents three binding sites, CATGTG, CATGTG, and CACATG, located at -2500 relative to TSS. R-box represents three binding sites, ATGGGAG, ATGGGAG, and TTCCCAC, located at +7900 relative to TSS. Lower panel: RBP-Jk and MITF protein-DNA complex on *TRPM1* promoter.

Supplementary Figure. 5: Phylogenetic profiling of meiotic specific genes.

a. Phylogenetic profiling of meiotic genes



b. gene expression in mouse tissues



c. DNA sequences of primers for uncharacterized genes

Gene symbol	5´-3´- DNA sequence	Amplified region
1700040L02Rik	GACGCCATCCTCTTGGACTA	394 bp
	CGTTGGCAATGGCTCTATG	
4930503L19Rik	CCCACAGCTCGTTTGTGTAT	304 bp
	CCATCCGCTGGAAGTCTTAAC	
Cenpw	GCCCTCCACCACAGTCACTA	153 bp
	GCCAATCGCTGAATAAAAAGTA	
2010007H12Rik	GGCTTGGACTTGAGCCTTAG	394
	GGTCTCGGTAGAGTCGCTTAGA	
Ccdc105	GACCATGACTTTGGGGCTAAT	322 bp
	AGTCCACATCGTGCCCTATC	

a) Meiosis-specific genes (synaptonemal complex genes: *Sycp2*, *Syce1*, *Syce2*, and *Syce3*, in bold) clustered based on their coevolution patterns across eukaryotic phyla with uncharacterized genes. Candidate meiotic-specific genes that were not annotated to function in meiosis or any other cellular process and were testes or ovary specific are indicated by red bold font. **b)** Expression patterns of uncharacterized genes in cDNA libraries from several mouse tissues. **c)** DNA sequences of primers used to amplify mRNA from genes in panel b.

Supplementary Table 1: **The phylogenetic profile database**. Each entry is Pab/Paa , when Pab is the best Blastp bit-score between a human gene 'a' and the top hit in organism 'b', Paa is the blast of the 'a' gene against itself. The result is a table of 19,017 genes by 86 species. ***See excel file***

Supplementary Table 2: *p*-values of MSigDB gene groups for the NPP with 86, 64, 43 and, 22 organisms and for BPP with 86 organisms. The columns represent the data for MSigDB groups and the calculated by *p*-values. The *p*-values were calculated by generating 1,000,000 random sets at the same size as the curated MSigDB set, and Co10 scores were calculated for all the random sets. A *p*-value < $1x10^{-6}$ was obtained if the score of the curated sets was higher than any of the 1,000,000 random sets with

similar size. To avoid biases, the random sets were generated only from the genes found in the MSigDB database.

			Number and percentage of significant co-evolved MSigDB groups in different methods																	
			q-values < 0.05										p-value < 10-6									
Collections	Subcollections	# of groups	NPP (86 Organisms)		NPP (64 Organisms)		NPP (43 Organisms)		NPP (22 Organisms)		BPP (86 Organisms)		NPP (86 Organisms)		NPP (64 Organisms)		NPP (43 Organisms)		NPP (22 Organisms)		BPP (86 Organisms)	
C1: positional gene sets	positional gene sets	320	10	3%	9	3%	6	2%	4	1%	4	1%	2	1%	3	1%	1	0%	1	0%	0	0%
C2: curated gene sets	chemical and genetic perturbations	2365	250	11%	208	9%	193	8%	159	7%	33	1%	47	2%	40	2%	43	2%	34	1%	9	0%
	BioCarta	217	41	19%	34	16%	27	12%	7	3%	8	4%	2	1%	3	1%	3	1%	1	0%	3	1%
	KEGG	186	98	53%	93	50%	85	46%	61	33%	22	12%	38	20%	34	18%	31	17%	24	13%	6	3%
	REACTOME	430	163	38%	159	37%	154	36%	110	26%	58	13%	67	16%	71	17%	60	14%	63	15%	23	5%
C3: motif gene sets	microRNA targets	219	54	25%	41	19%	43	20%	20	9%	0	0%	4	2%	5	2%	1	0%	1	0%	0	0%
	transcription factor targets	584	143	24%	135	23%	89	15%	54	9%	5	1%	6	1%	6	1%	2	0%	4	1%	0	0%
C4: computational gene sets	cancer modules	443	81	18%	77	17%	71	16%	76	17%	33	7%	25	6%	23	5%	23	5%	27	6%	8	2%
	cancer gene neighborhoods	427	149	35%	132	31%	141	33%	108	25%	22	5%	48	11%	46	11%	51	12%	48	11%	9	2%
C5: GO gene sets	GO biological process	793	136	17%	121	15%	100	13%	64	8%	42	5%	38	5%	28	4%	30	4%	16	2%	7	1%
	GO cellular component	215	50	23%	41	19%	43	20%	40	19%	9	4%	18	8%	18	8%	19	9%	8	4%	3	1%
	GO molecular function	395	102	26%	99	25%	93	24%	70	18%	38	10%	45	11%	37	9%	40	10%	21	5%	8	2%
	TOTAL	6594	1277	19%	1149	17%	1045	16%	773	12%	274	4%	340	5%	314	5%	304	5%	248	4%	76	1%

See excel for the full data

Supplementary Table 3: Co10 scores, *p*-values, and *q*-values of HPO gene groups. The columns represent the data for HPO groups and the calculated by *p*-values and *q*-values. The *p*-values were calculated by generating 1,000,000 random sets at the same size as the curated HPO set, and Co10 scores were calculated for all the random sets. A *p*-value < $1x10^{-6}$ was obtained if the score of the curated sets was higher than any of the 1,000,000 random sets with similar size. To avoid biases, the random sets were generated only from the genes found in the HPO database. ***See excel file***

Supplementary Table 4: The overlap between the genes, which contribute to the Co10 score in pairs of Significant HPOs. The overlap between the HPOs was used to generate the HPO network (Figure. 4). ***See excel file***

Supplementary Table 5: The mapping of the human genes into coevolved clusters. The 1076 phylogenetically clustered human genes, and the average phylogenetic profile of the genes in the cluster. Column A - the index identifier of the cluster. Column B – the list of genes in the cluster. Column C - the number of genes in the cluster. Column D to CK - the average phylogenetic protein conservation of the genes in the cluster across 86 genomes. ***See excel file***

Supplementary Table 6: Gene overlap between coevolved clusters and functional groups. The functional gene groups and disease phenotypic groups (columns A-B), their size (column C) and the overlap statistic (column F to H) between these groups and the 1076 phylogenetic clusters (column D) that indexes in Supplementary Table 5. To

eliminate the effect of homologous gene families, the over-representation p-values and FDR (columns L-M) were also calculated using the number of protein families in the functional groups (column I), disease phenotypic groups (column j), and in the overlap (column K) instated of all the genes in each group. ***See excel file***

Supplementary Table 7: Phylogenetic profile of the 50 genes that are mostly correlated with MITF. The correlation coefficients were calculated using the normalized phylogenetic profile matrix and the genes are rank ordered. Each row represents a gene and each entry is Pab/Paa , when Pab is the best Blastp bit-score between a human gene 'a' and the top hit in organism 'b', Paa is the blast of the 'a' gene against itself. ***See excel file***

Supplementary Table 8: Presence of RBP-Jk binding sites in MITF target gene promoters.

Gene Name	RBP-Jk BS
PMEL	Yes
TYRP1	No
TYR	Yes
TBX2	Yes
C-MET	Yes
MART1	No
DCT	Yes
CDKN1A	Yes
CDK2	Yes
BCL2	No
AIM1	Yes
DICER1	Yes
TRPM1	Yes

Supplementary Table 9: Primer and siRNA sequences.

siRNA

reagent	**sequence**		**company**
si control	sense	5' AAUUCUCCGAACGUGUCACGU 3'	ABI Inc.
	antisense	5' ACGUGACACGUUCGGAGAAUU 3'	
si MITF #1	sense	5' GGCUUUCUAGAAAGAAUAA 3'	ABI Inc.
	antisense	5' UUAUUCUUUCUAGAAAGCC 3'	
si MITF #2	sense	5' GGUGAAUCGGAUCAUCAAG 3'	designed by Carreira et al, *Nature* 2005
	antisense	5' CUUGAUGAUCCGAUUCACC 3'	and purchased from ABI Inc.
TriFECTa siRBPJk	duplex 1	5' GGCUGGAAUACAAGUUGAACAAACT 3' 3' CACCGACCUUAUGUUCAACUUGUUUGA 5'	IDT
	duplex 2	5' CCAAGGAACUUGUAUAUUGUAUAAG 3' 3' CUGGUUCCUUGAACAUAUAACAUAUUC 5'	
	duplex 3	5' CACAGAUAAGGCAGAGUAUACAUTT 3' 3' UCGUGUCUAUUCCGUCUCAUAUGUAAA 5'	

Primers - RT-PCR

primer name	sequance	
h-mMITF	Forward	5' CATTGTTATGCTGGAAATGCTAGAA 3'
	Reverse	5' GGCTTGCTGTATGTGGTACTTGG 3'
hActin	Forward	5' ATTGCCGACAGGATGCAGAA 3'
	Reverse	5' GCTGATCCACATCTGCTGGAA 3'
PZLF2	Forward	5' ATGGTCCGGCTCTCTGACTTC 3'
	Reverse	5' TGAAGACGGAGATGATGCAGG 3'
Hes5	Forward	5' GTCAACACGACACCGGATAAA 3'
	Reverse	5' TCAGCTGGCTCAGACTTTCA 3'
TRPM1	Forward	5' TCTTGAATTCCAGGGTGGCGGATA 3'
	Reverse	5' GTTCCAGCTGCCAATCTTTCACCA 3'

Primers - ChIP

primer name	sequance	
HES1 promtor	Reverse	5' GGACACACACACACACACACCC 3'
		5' CCGGGAGAAGCCAAGAAGGTAAAT 3'
TPRM1 pro Ebox1	Forward	5' ACAGCAAATCCAACAGAGCTCCCA 3'
	Reverse	5' ACAGCAAATCCAACAGAGCTCCCA 3'
TPRM1 pro Ebox2	Forward	5' GTAATTAGTGCCATGTGCCGCCTT 3'
	Reverse	5' TGGCTGGAAATGTCAGCAGGGTTT 3'
TPRM1 pro Ebox3	Forward	5' CATCGCTTCACAGCAATCATGAGG 3'
	Reverse	5' ATCTGAGCTTACCTGCCACAGCA 3'
TRPM1 int1_A	Forward	5' TGTGTGTAAATGCTGGACCTGGGA 3'
	Reverse	5' ACTGAGGGCAACATGTCTCTTGCT 3'
TRPM1 int1_B	Forward	5' AATGCTGGACCTGGGAAGATGGAA 3'
	Reverse	5' ACTGAGGGCAACATGTCTCTTGCT 3'
TRPM1pro Rbox	Forward	5' AGCAAAGGACAGAGAACATCGGGA 3'
	Reverse	5' CAGCAGTTGGGAAGGTGTTGTGTT 3'
TRPM1 ex2	Forward	5' CGGGAATGTATCTTTGTAATTCCTAGC 3'
	Reverse	5' TCTGGACTCCTCTTTCTGCCTCTT 3'
TRPM1 ex4	Forward	5' CAGGCAGTTCTTGCTTGGACATGA 3'
	Reverse	5' GTTCCAGCTGCCAATCTTTCACCA 3'
TRPM1 ex5	Forward	5' TTGTGCTTGCTTCCCTTGTTGGTC 3'
	Reverse	5' TGTGGGAGTTGTTGAGCACAGAGA 3'
TRPM1 ex9	Forward	5' TGTGTTCAGAATGGGTTCTGAGGG 3'
	Reverse	5' TCTAGCACCAGCAGACATAGTCCA 3'
bACTpromotor	Forward	5' CCCACCTCACCACTCTTCCTATTT 3'
	Reverse	5' TCAGCCTAGAGGAACCTGCCTT 3'
bACTpromotor2	Forward	5' GGAGTGTGGTCCTGCGACTTCTAA 3'
	Reverse	5' TAGCTAAATGTGCTGGGTGGGTCA 3'
bACT ex2	Forward	5' TCACCATGGATGATGATATCGCCG 3'
	Reverse	5' TCCTGTGCAGAGAAAGCGCC 3'
bACT ex4	Forward	5' AGCCGTGTTCTTTGCACTTTCTGC 3'
	Reverse	5' TAGCACAGCCTGGATAGCAACGTA 3'
bACT ex6	Forward	5' AATGTGGCCGAGGACTTTGATTGC 3'
	Reverse	5' AGGATGGCAAGGGACTTCCTGTAA 3'

Example of particular intersection of MsigDB and HPO clusters with phylogenetic clusters

P-value of overlap between the HPO Acidosis and phylogenetic cluster number 305 was calculated using hyper-geometric test.