CYBRD1 as a modifier gene that modulates iron phenotype in *HFE* p.C282Y homozygous patients

Sara Pelucchi,[1] Raffaella Mariani,[2] Stefano Calza,[3,4] Anna Ludovica Fracanzani,[5] Giulia Litta Modignani,[1] Francesca Bertola,[6] Fabiana Busti,[7] Paola Trombini,[2] Mirella Fraquelli,[8] Gian Luca Forni,[9] Domenico Girelli,[7] Silvia Fargion,[5] Claudia Specchia,[3] and Alberto Piperno[1,2,6]

[1]Department of Clinical Medicine and Prevention, University of Milano-Bicocca, Monza, Italy; [2]Center for Diagnosis and Treatment of Hemochromatosis, S. Gerardo Hospital, Monza, Italy; [3]Department of Biomedical Sciences and Biotechnologies, University of Brescia, Brescia, Italy; [4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [5]Department of Medicine, IRCCS, Ospedale Maggiore Policlinico, University of Milano, Milano, Italy; [6]Consortium for Human Molecular Genetics, Monza, Italy; [7]Department of Medicine Policlinico GB Rossi, University of Verona, Verona, Italy; [8]Second Division of Gastroenterology, IRCCS, Ospedale Maggiore Policlinico, University of Milano, Milano, Italy; and [9]Centre of Microcitemia and Congenital Anemias, Galliera Hospital, Genova, Italy

Citation: Pelucchi S, Mariani R, Calza S, Fracanzani AL, Modignani GL, Bertola F, Busti F, Trombini P, Fraquelli M, Forni GL, Girelli D, Fargion S, Specchia C, and Piperno A. CYBRD1 *as a modifier gene that modulates iron phenotype in* HFE *p.C282Y homozygous patients. Haematologica 2012;97(12):1818-1825. doi:10.3324/haematol.2012.062661*

Online Supplementary Design and Methods

Selection of single nucleotide polymorphisms

Illumina assigns designability ranks and single nucleotide polymorphism (SNP) scores. They are items strictly linked since they both give the same information about the ability to design a successful assay. Illumina SNP scores range from 0 to 1.1 and designability rank is represented by scores 0, 0.5 or 1. A SNP_Score < 0.4, corresponding to a designability rank of 0, gives a low success rate with consequent high risk of performing an oligo pull assay (OPA). A SNP_Score from 0.4 to 0.6 corresponds to a designability rank of 0.5 and gives a moderate success rate with a consequent moderate risk of performing an OPA. A SNP_Score from 0.6 to 1.1, equivalent to a designability rank of 1, gives a high success rate with a consequent low risk of performing an OPA. All the SNPs we analyzed had Illumina SNP scores ≥1, a designability rank ≥1, no failure code, validation class=3 and GoldenGate® validation. Validation class data and validation bin are additional items to consider; they are numeric and textual representations of genotyping reaction feasibility. GoldenGate® validation bin SNPs have a validation class of 3 and SNP_scores of 1.1. Two-hit validated SNPs have a validation class of 2 (meaning that both alleles of the SNP have been read by two different methods and in two independent populations) and SNP_scores from 0 to 1. Not validated SNPs have a validation class of 1 and a SNP_Score of 0. We analyzed 133 GoldenGate® and 78 two-hit validated class SNPs, with 3 and 2 validation bin, respectively. SNP scores were 1.1 for GoldenGate® ones and between 0.668 and 0.997 for two-hit validated SNPs (mean 0.886).

A set of non-redundant tag SNPs was identified for each region, so that all the SNPs with a minor-allele frequency ≥0.05 in the database had a pairwise r^2≥0.80 (*www.hapmap.org - hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap3r2_B36/#search*). Tagging was performed using the algorithm implemented in Tagger. The linkage disequilibrium blocks were determined using data from HapMap data release #28, on National Center for Biotechnology (NCBI) B36 assembly, dbSNP b126.

Among all the TagSNPs selected from Haploview we retained only those belonging to coding, intronic, 5′ and 3′ untranslated regions with similar proportions in order to cover the full gene.

Genotyping of single nucleotide polymorphisms

For laboratory quality assurance, we qualified SNPs that had Illumina GenCall_10 scores ≥0.4 and call rates ≥88%. So, we excluded SNPs with GenCall_10 scores below 0.40 and/or call rates below 88%. VeraCode Raw data generated from the genotyping were analyzed by the GenomeStudio software to define, for each SNP, the called genotypes into the three different cluster areas. Data were then processed in order to infer all SNP genotypes via a genotyping cluster. All genotypes were manually checked and re-scored if any errors in calling homozygous or heterozygous clusters were evident. Samples falling out of these cluster areas corresponding to the different genotypes were failed. Four duplicate samples were genotyped for all assays for quality control with 100% reproducibility. All SNPs showed high-genotyping quality: the genotyping call rate for the studied SNPs was in the range of 99-100%.

Statistical methods

(i)Principal component analysis

Principal component analysis (PCA) is one of the most used and valuable applications of linear algebra, being a simple, non-parametric method of dimensionality reduction, i.e. of extracting relevant information from a complex set of data.

The procedure allows to extract, from a number p of correlated variables, as many as p new factors derived as linear combinations of the original variables. The biggest advantage is that one or few of the PCA factors accounts for a great proportion of the total variance (hence information) of the p variables. Let us consider only two variables with approximately the same variance and reasonably high correlation. Then let us plot them in a scatter-plot. The PCA would draw two new *orthogonal* axes, one passing along the direction with higher variance in the cloud of points and the second being perpendicular to the first one. These axes are the new coordinates for the derived PCA factors, and can be simply computed as a linear combination of the original

variables. In the ideal setting of two variables with very high correlation, the new first component (axis) would account for the large majority of information (variability) contained in the data, and could be used as a surrogate of the two variables combined. The same idea can be applied to any number *p* of variables.

The coefficients used to compute the different linear combinations are traditionally called *loadings*. The proportion of variance explained or accounted for by the *p* new axis is an indication of the information loss by the dimensionality reduction.

(ii) Classification

A classifier is defined as a model describing the specific classification algorithm, such as support vector machine (SVM), K-nearest neighbors (KNN), etc. Every classifier was built iteratively on all samples but a single one (the so-called "*out of the bag*" sample). Specifically, for every classification algorithm we fitted 138 different models (where 138 is the overall number of patients used), each one built on 137 samples excluding each sample in turn. In every single iteration we used the model fitted on 137 samples to predict the status (case/control, i.e. high/low tertile of serum ferritin, iron removed and transferrin saturation combined values) of the excluded (*out of the bag*) sample based on the covariates in the model (sex, age, alcohol and a given number of SNP). In iteration one we would then exclude sample one and build the classifier on the remaining 137 samples; such a classifier is then used to predict the status for sample one. The same holds true for sample two and so on.

Online Supplementary Table S1. Two hundred and fourteen single nucleotide polymorphisms analyzed in 50 genes.

SNPs	Chromosome	Gene name	Location	Function
rs4693924	4	ABCG2	Intron	Superfamily of ATP-binding cassette (ABC) transporters
rs2054576	4	ABCG2	Intron	
rs2725256	4	ABCG2	Intron	
rs4148155	4	ABCG2	Intron	
rs2622624	4	ABCG2	Intron	
rs1901531	15	B2M	Intron	HFE pathway
rs1801621	11	BEST1	3'UTR	Carrier calcium-activated chloride-ion channels
rs6077060	20	BMP2	5' UTR	BMP pathway
rs235768	20	BMP2	Coding	
rs3178250	20	BMP2	3'UTR	
rs6054512	20	BMP2	3'UTR	
rs173107	20	BMP2	3'UTR	
rs235756	20	BMP2	3'UTR	
rs910141	20	BMP2	3'UTR	
rs17563	14	BMP4	Coding	BMP pathway
rs762642	14	BMP4	Intron	
rs4901474	14	BMP4	3'UTR	
rs3812163	6	BMP6	5UTR	BMP pathway
rs6910759	6	BMP6	Intron	
rs267201	6	BMP6	Intron	
rs1225934	6	BMP6	Intron	
rs1044104	6	BMP6	3'UTR	
rs9325886	10	BMP9	3'UTR	BMP pathway
rs9971293	10	BMP9	Intron	
rs11204215	10	BMP9	5'UTR	
rs3905377	10	BMPR1A	5'UTR	BMP pathway
rs2354353	10	BMPR1A	Intron	
rs2883420	10	BMPR1A	Intron	
rs10887666	10	BMPR1A	Intron	
rs7091555	10	BMPR1A	Intron	
rs7074064	10	BMPR1A	Intron	
rs4401458	4	BMPR1B	Intron	BMP pathway
rs7661049	4	BMPR1B	Intron	
rs6815044	4	BMPR1B	Intron	
rs9997720	4	BMPR1B	Intron	
rs3821964	4	BMPR1B	Intron	
rs3796443	4	BMPR1B	Intron	
rs11097457	4	BMPR1B	3'UTR	
rs13010656	2	BMPR2	Intron	BMP pathway
rs6751210	2	BMPR2	Intron	
rs7575056	2	BMPR2	Intron	
rs12467409	2	BMPR2	Intron	
rs1048829	2	BMPR2	3'UTR	
rs1053709	3	CP	Coding	Iron reductase
rs701754	3	CP	Intron	

continued on the next page

continued from the previous page

rs773050	3	CP	Intron	
rs701753	3	CP	Coding	
rs701748	3	CP	5'UTR	
rs3806562	2	CYBRD1	5'UTR	Iron reductase
rs3806566	2	CYBRD1	5'UTR	
rs884409	2	CYBRD1	5'UTR	
rs960748	2	CYBRD1	Intron	
rs17554	2	CYBRD1	Intron	
rs10455	2	CYBRD1	Coding	
rs2542938	2	CYBRD1	3'UTR	
rs1435166	1	EGLN1	Intron	Hypoxia pathway
rs2486742	1	EGLN1	Intron	
rs1538664	1	EGLN1	Intron	
rs7544596	1	EGLN1	5'UTR	
rs3736329	19	EGLN2	Intron	Hypoxia pathway
rs10405596	19	EGLN2	3'UTR	
rs1680710	14	EGLN3	3'UTR	Hypoxia pathway
rs1680694	14	EGLN3	Intron	
rs1047881	1	FLVCR1	5'UTR	Heme pathway
rs12756625	1	FLVCR1	Intron	
rs12125982	1	FLVCR1	Intron	
rs10779594	1	FLVCR1	Intron	
rs1390501	1	FLVCR1	Intron	
rs3207090	1	FLVCR1	Coding	
rs4932178	15	FURIN	5'UTR	Hepcidin cleavage
rs6227	15	FURIN	3'UTR	
rs12459782	19	GDF15	5'UTR	Growth factor
rs1059519	19	GDF15	Coding	
rs1227731	19	GDF15	Intron	
rs16982345	19	GDF15	3'UTR	
rs8101249	19	GDF15	3'UTR	
rs10405246	19	USF2	Intron	Transcription factor
rs1882694	19	USF2	3'UTR	
rs8101606	19	HAMP	Intron	Hepcidin pathway
rs7251432	19	HAMP	Intron	
rs12971321	19	HAMP	3'UTR	
rs1264218	X	HEPH	Intron	Iron oxidase
rs5919024	X	HEPH	Intron	
rs1800702	6	HFE	5'UTR	Hepcidin regulator
rs2794719	6	HFE	Intron	
rs9366637	6	HFE	Intron	
rs2858996	6	HFE	Intron	
rs707889	6	HFE	Intron	
rs16827043	1	HFE2	5'UTR	Hepcidin regulator
rs7536827	1	HFE2	5'UTR	
rs1535921	1	HFE2	3'UTR	
rs2301106	14	HIF1A	Intron	Hypoxia pathway
rs12434438	14	HIF1A	Intron	
rs10873142	14	HIF1A	Intron	
rs2301113	14	HIF1A	Intron	
rs2057482	14	HIF1A	3'UTR	

continued on the next page

continued from the previous page

rs11684885	2	HIF2A	Intron	Hypoxia pathway
rs11689011	2	HIF2A	Intron	
rs6756667	2	HIF2A	Intron	
rs1374748	2	HIF2A	Intron	
rs7571218	2	HIF2A	Intron	
rs13424253	2	HIF2A	3'UTR	
rs9924964	16	HP	5'UTR	Heme pathway
rs2856836	2	IL1A	3'UTR	Inflammation
rs3783546	2	IL1A	Intron	
rs1800587	2	IL1A	5'UTR	
rs1878319	2	IL1A	5'UTR	
rs2069832	7	IL6	Intron	Inflammation
rs2069849	7	IL6	Coding	
rs4601580	1	IL6R	Intron	Inflammation
rs7518199	1	IL6R	Intron	
rs4553185	1	IL6R	Intron	
rs4845625	1	IL6R	Intron	
rs4129267	1	IL6R	Intron	
rs11265618	1	IL6R	Intron	
rs4072391	1	IL6R	3'UTR	
rs4297112	9	IRP1	Intron	Cell iron regulation
rs7874815	9	IRP1	Intron	
rs10970971	9	IRP1	Intron	
rs10813813	9	IRP1	Intron	
rs3780474	9	IRP1	Intron	
rs4878497	9	IRP1	Intron	
rs10813816	9	IRP1	Intron	
rs10970978	9	IRP1	Intron	
rs7042042	9	IRP1	3'UTR	
rs17483548	15	IRP2	5'UTR	Cell iron regulation
rs12916396	15	IRP2	Intron	
rs2938674	15	IRP2	Intron	
rs13180	15	IRP2	Coding	
rs2292116	15	IRP2	Intron	
rs16969906	15	IRP2	3'UTR	
rs3814526	9	LCN2	5'UTR	Iron Carrier
rs721183	8	MFRN1	5'UTR	Mitochondrial iron transporter
rs4872154	8	MFRN1	Intron	
rs1047384	8	MFRN1	Coding	
rs922516	15	NEO1	Intron	HJV pathway
rs1979409	15	NEO1	Intron	
rs3736510	15	NEO1	Coding	
rs2292915	15	NEO1	Intron	
rs1878940	15	NEO1	3'UTR	
rs739439	17	SARM1	3'UTR	Inflammation
rs149411	12	DMT1	3'UTR	Iron absorption
rs161047	12	DMT1	Intron	
rs445520	12	DMT1	Intron	
rs364627	12	DMT1	Intron	
rs870843	3	SLC25A38	Intron	Mitochondrial carrier

continued on the next page

continued from the previous page

rs6890	3	SLC25A38	3'UTR	
rs2352262	2	SLC40A1	3'UTR	Iron exporter
rs2304704	2	SLC40A1	Coding	
rs4145237	2	SLC40A1	Intron	
rs1439812	2	SLC40A1	Intron	
rs3811621	2	SLC40A1	5'UTR	
rs6537355	4	SMAD1	5'UTR	Signal transduction
rs2289737	4	SMAD1	Intron	
rs714195	4	SMAD1	Intron	
rs11724813	4	SMAD1	Intron	
rs12457540	18	SMAD4	Intron	Signal transduction
rs2276163	18	SMAD4	Intron	
rs8084630	18	SMAD4	Intron	
rs8096092	18	SMAD4	Intron	
rs948588	18	SMAD4	Intron	
rs9304407	18	SMAD4	Intron	
rs6596289	5	SMAD5	Intron	Signal transduction
rs13179769	5	SMAD5	Intron	
rs10068371	5	SMAD5	Intron	
rs10515478	5	SMAD5	Intron	
rs7031	5	SMAD5	3'UTR	
rs17804636	13	SMAD8	3'UTR	Signal transduction
rs7993661	13	SMAD8	Intron	
rs9547689	13	SMAD8	Intron	
rs9576129	13	SMAD8	5'UTR	
rs1053005	17	STAT3	3'UTR	Inflammation
rs3744483	17	STAT3	3'UTR	
rs8074524	17	STAT3	Intron	
rs6503695	17	STAT3	Intron	
rs1026916	17	STAT3	Intron	
rs17405722	17	STAT3	5'UTR	
rs838082	2	STEAP3	Intron	Iron reductase
rs1867749	2	STEAP3	Intron	
rs3731603	2	STEAP3	3'UTR	
rs1530561	2	STEAP3	3'UTR	
rs8177178	3	TF	5'UTR	Iron transport
rs8177213	3	TF	Intron	
rs8177240	3	TF	Intron	
rs3811647	3	TF	Intron	
rs1525889	3	TF	Intron	
rs10247962	7	TFR2	Intron	Hepcidin regulator
rs2075674	7	TFR2	Coding	
rs7457868	7	TFR2	Intron	
rs4727457	7	TFR2	Intron	
rs4434553	7	TFR2	3'UTR	
rs1052897	7	TFR2	3'UTR	
rs6772320	3	TFRC	3'UTR	Iron uptake
rs3326	3	TFRC	Intron	
rs3827556	3	TFRC	Intron	
rs3817672	3	TFRC	Coding	
rs4820268	22	TMPRSS6	Coding	HJV pathway
rs2543519	22	TMPRSS6	Intron	
rs2179229	22	TMPRSS6	Intron	
rs2235323	22	TMPRSS6	Intron	
rs2235324	22	TMPRSS6	Coding	
rs2743824	22	TMPRSS6	Intron	
rs732755	22	TMPRSS6	Intron	
rs855791	22	TMPRSS6	Coding	
rs228910	22	TMPRSS6	5'UTR	
rs779805	3	VHL	5'UTR	Hypoxia pathway
rs1642742	3	VHL	3'UTR	
rs17610448	3	VHL	3'UTR	
rs12544577	8	ZIP14	5'UTR	Metal ion transporter
rs11136017	8	ZIP14	Intron	
rs2280521	8	ZIP14	Intron	
rs12545575	8	ZIP14	Intron	
rs10101909	8	ZIP14	Intron	
rs12679702	8	ZIP14	3'UTR	