Research article

SNP selection for genes of iron metabolism in a study of genetic modifiers of hemochromatosis

Clare C Constantine*[1,2], Lyle C Gurrin[1], Christine E McLaren[2], Melanie Bahlo[3], Gregory J Anderson[4], Chris D Vulpe[5], Susan M Forrest[6], Katrina J Allen[7], Dorota M Gertig[1] and the HealthIron Investigators

Address: [1]The Centre for Molecular, Environmental, Genetic and Analytic (MEGA) Epidemiology, School of Population Health, The University of Melbourne, Melbourne, Australia, [2]Department of Epidemiology, University of California, Irvine, USA, [3]Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, [4]Iron Metabolism Laboratory, The Queensland Institute of Medical Research, Brisbane, Australia, [5]Department of Nutritional Sciences and Toxicology, University of California, Berkeley, USA, [6]Australian Genome Research Facility, Melbourne, Australia and [7]Murdoch Childrens Research Institute, Melbourne, Australia

Email: Clare C Constantine* - ccconsta@uci.edu; Lyle C Gurrin - lgurrin@unimelb.edu.au; Christine E McLaren - cmclaren@uci.edu; Melanie Bahlo - bahlo@wehi.edu.au; Gregory J Anderson - Greg.Anderson@qimr.edu.au; Chris D Vulpe - vulpe@berkeley.edu; Susan M Forrest - sue.forrest@agrf.org.au; Katrina J Allen - katie.allen@rch.org.au; Dorota M Gertig - dgertig@vccr.org; the HealthIron Investigators - ccconsta@uci.edu

* Corresponding author

Abstract

Background: We report our experience of selecting tag SNPs in 35 genes involved in iron metabolism in a cohort study seeking to discover genetic modifiers of hereditary hemochromatosis.

Methods: We combined our own and publicly available resequencing data with HapMap to maximise our coverage to select 384 SNPs in candidate genes suitable for typing on the Illumina platform.

Results: Validation/design scores above 0.6 were not strongly correlated with SNP performance as estimated by Gentrain score. We contrasted results from two tag SNP selection algorithms, LDselect and Tagger. Varying r^2 from 0.5 to 1.0 produced a near linear correlation with the number of tag SNPs required. We examined the pattern of linkage disequilibrium of three levels of resequencing coverage for the transferrin gene and found HapMap phase 1 tag SNPs capture 45% of the $\geq$ 3% MAF SNPs found in SeattleSNPs where there is nearly complete resequencing. Resequencing can reveal adjacent SNPs (within 60 bp) which may affect assay performance. We report the number of SNPs present within the region of six of our larger candidate genes, for different versions of stock genotyping assays.

Conclusion: A candidate gene approach should seek to maximise coverage, and this can be improved by adding to HapMap data any available sequencing data. Tag SNP software must be fast and flexible to data changes, since tag SNP selection involves iteration as investigators seek to satisfy the competing demands of coverage within and between populations, and typability on the technology platform chosen.

Background

Single nucleotide polymorphisms (SNPs), changes in a single base pair of the DNA sequence, are the most frequently occurring form of variation in the human genome. Many genes have a large number of SNPs, and it is acknowledged that there are more than 10 million SNPs across the human genome, making it impossible for cost-effective genotyping of all of them in studies of disease, even in very small samples. We can, however, reduce the genotyping burden by exploiting the strong correlation between some SNPs that are close together on the genome. This is due to the phenomenon of linkage disequilibrium (LD), or non-random association of SNP alleles at the population level, due to the sharing by multiple individuals of ancestral chromosomal segments. These segments, or haplotypes, are combinations of particular SNP alleles on the same chromosome that tend to segregate together. By choosing a subset of maximally informative SNPs, or "tag" SNPs, to represent these haplotypes, the number of SNPs to be genotyped in a larger sample can be reduced without losing the ability to capture most of the variation, and in particular any association between unmeasured "causal" alleles and the disease outcome measured on individuals in the sample. This is an increasingly common approach to genetic association studies, since it reduces costs but retains much of the information about linkage disequilibrium patterns across the human genome. It is the underlying principle behind the Hap-Map [1] project. An additional assumption behind this approach is the idea of the common variant common disease hypothesis. It is assumed that the variant, and hence its haplotype, are relatively common in the general population and hence will be ascertained using this approach. If, however, the disease is caused by a rare variant, this approach may fail to detect association.

Many researchers are now turning to large, publicly available databases of SNPs, which provide a catalogue of human genetic variation, in order to choose a set of informative "tag" SNPs for genotyping in association studies of disease. The International HapMap Project [2] and the Seattle SNPs project are complemented by several NIH initiatives including the National Institute of Environmental Health Sciences Environmental Genome Project (NIEHS EGP) and the National Heart Lung and Blood Institute Resequencing and Genotyping Project. The choice of tag SNPs is made more challenging when study subjects are from multiple populations, since the transferability of tag SNPs depends on similarity of linkage disequilibrium patterns. It is also desirable to incorporate resequencing data from local case and control samples generated during a "SNP discovery" phase. A further complication is the need for the chosen tag SNPs to have high probability of successfully being genotyped on

the high-throughput platform being used to process the samples.

Iron is essential for life and consequently body iron levels are tightly regulated in humans. There are disease states associated with having either too little (iron deficiency) or too much (iron overload) iron, the former usually due to inadequate dietary iron or excessive iron loss and the latter usually associated with mutations in proteins that regulate intestinal iron absorption. About 90% of clinical cases of iron overload (hemochromatosis) in populations of northern European origin are homozygous for the 845 G $\rightarrow$ A mutation in the *HFE* gene responsible for the C282Y substitution in the HFE protein [3]. Affected individuals are characterised by high transferrin saturation (a measure of the amount of circulating iron), an increased serum ferritin (a measure of iron storage) and the associated clinical symptoms of iron overload (fatigue, arthritis, abnormal liver function and ultimately permanent tissue damage). Despite the high prevalence of mutations in the *HFE* gene, the phenotypic expression of hemochromatosis varies considerably and both environmental and genetic factors appear to make important contributions to this variation. The *HealthIron* Study seeks to find associations in Caucasian subjects between common polymorphisms in candidate genes of iron metabolism and variations in the iron phenotype in HFE-associated hemochromatosis.

Here we describe the methods used to select SNPs for genotyping of samples from the HealthIron Study on the basis of sequence data available from a number of different sources, both public and proprietary, with particular reference to iron homeostasis genes. The selected SNPs were subsequently genotyped using the Illumina platform which genotypes one to four multiples of 384 SNPs in a multiplex reaction. A goal of this research was to identify 384 SNPs based on a consideration of 35 genes involved in iron metabolism.

Methods

We used four data sources in order to obtain thorough SNP coverage of the genes and to allow detection of either direct association between iron phenotypes and a causal variant or indirect association with a marker that is in linkage disequilibrium (LD) with a causal variant. We used a minor allele frequency (MAF) cut-off of $\geq 3\%$ and r^2 (the square of the correlation between SNPs) of 0.8. The MAF cut-off of $\geq 3\%$ was chosen as a compromise between power (i.e. sample size, 3% represents 6 heterozygotes among 94 individuals sequenced) and detection of rare alleles with large effect.

The four data sets used were:

1. HealthIron: Resequencing of selected candidate genes has been performed on two groups of individuals of northern European descent (94 C282Y homozygotes and 94 chosen randomly without regard to *HFE* genotype) from a large population-based cohort study [4]. We report here only on the randomly chosen individuals who were resequenced for six genes.

2. HEIRS ancillary study NHLBI: The Hemochromatosis and Iron Overload Screening [5] (HEIRS), ancillary study on iron deficiency carried out initial SNP identification from 14 candidate genes by resequencing of five populations: 24 African American, 24 Yorubans (HapMap YRI), 47 Caucasian with north-west European ancestry (HapMap CEU), 48 Hispanic descent, 45 Chinese (HapMap CHB). This resequencing was completed by the Resequencing and Genotyping Service of the National Heart Lung and Blood Institute (NHLBI RS&G). Accession numbers are TFR2: DQ496110, TF: DQ525716, HFE2: DQ309445, HCP1: DQ496103, TFRC: DQ496099, PGRMC2: DQ496105, PGRMC1: DQ496104, IREB2: DQ496102, HEPH: DQ496100, HAMP: DQ496109, FTH1: DQ496108, FLVCR: DQ496107, CYBRD1: DQ496101, and ACO1: DQ496106.

3. SeattleSNPs [6]: This is funded as part of the NHLBI Programs for Genomic Applications (PGA). It aims to investigate the associations between SNPs in candidate genes and pathways that underlie inflammatory responses in humans. Individual investigators can nominate candidate genes to be resequenced for SNP discovery. Three of the genes of iron metabolism targeted by the HealthIron and HEIRS ancillary projects have been resequenced by SeattleSNPs (TF, HMOX1, TNFα all panel 1: 23 CEPH (9 are HapMap children), 24 African Americans – which are the same as in the HEIRS ancillary study).

4. HapMap (publicly available [7]): The International HapMap Project [2] is analyzing DNA from populations with African, Asian, and European ancestry to generate a catalogue of common SNPs across the whole genome in humans.

Tag SNP selection programs
There are many algorithms and software packages designed to select tag SNPs from large arrays of genotype data. We review briefly the two packages that we used to select a subset of SNPs for further genotyping in a larger sample of individuals from our cohort study.

LDSelect
The *LDselect* algorithm [8,9] partitions the SNPs into "bins", that is, each SNP is a member of one and only one bin. In a given bin there is at least one SNP that has a pairwise r^2 exceeding a user-specific threshold (*e.g.* 80%) with

each of the other SNPs in that same bin, where r^2 is the correlation between two SNPs calculated using the genotype data available from one of the data sources. There may be several such tag SNPs in each bin, but not all SNPs in each bin are necessarily tag SNPs for that bin; it is possible that some pairs of SNPs in the same bin have pairwise r^2 values that do not exceed the threshold. Each of the tag SNPs in a particular bin can be used to represent the allelic variation of SNPs within each bin, and is a candidate for genotyping in a larger sample.

Tagger in Haploview
The *Tagger* software begins by using a tagging algorithm similar to LDselect, where SNPs are captured by requiring pairwise association with at least one of a series of single tagging markers at a prescribed threshold value of the pairwise correlation r^2 (software available as part of Haploview [10]). Tagger then seeks to further reduce the number of tag SNPs by attempting to replace each tag with a multi-marker predictor based on the remaining tag SNPs. Any proposed multi-marker combination is checked to ensure that it can capture the SNPs originally represented by the replaced tag SNP with a value of r^2 exceeding the prescribed threshold, and if not, the original tag SNP is retained. Further details can be found in de Bakker et al. [11] and Barrett et al. [10].

Challenge with Illumina platform – validation scores
The Illumina corporation use an algorithm (accessed via a service provided free of charge to prospective clients) to generate a validation score for a specified SNP. The validation score, which takes values between 0 and 1, is calculated from the 200 base pair genetic sequence surrounding each SNP. It is an estimate of the likelihood that an assay for that SNP will work successfully, i.e. genotype most individuals accurately. If the SNP has previously been successfully genotyped on the Illumina platform the SNP is given a validation score of 1.1 and it becomes a so-called "Golden Gate" SNP. Illumina recommends not using SNPs with validation scores below 0.6, and including additional SNPs for redundancy to overcome loss of information due to SNP failure even for SNPs with high validation scores. SNPs within 60 bp of each other and tri-allelic SNPs cannot be typed.

We attempted to satisfy the requirement of genotyping only SNPs with high validation scores and to incorporate redundancy by (i) selecting tag SNPs separately from each dataset where sequence data for a gene was available from more than one data source and (ii) selecting additional tag SNPs to be genotyped in large bins in case the first tag SNP failed. The selection of high validation score SNPs should increase the genotyping rate and minimise the chance of assay failure. It is straightforward to incorporate these rules into the selection process when using LDselect

which lists all the alternative tag SNPs for each bin. There-fore it is necessary to run the program only once and choose the highest validation score SNP from the nominated alternative tag SNPs (provided at least one tag SNP has validation score > 0.6), since changing the selection of one or more tag SNPs within a bin does not affect the selection process of tag SNPs in the other bins. We also took advantage of the alternate tagSNP listing from LDselect to choose a second redundant tagSNP for large bins in case the first tagSNP assay failed.

In contrast, this is not possible using Tagger, although Tagger now has the ability to exclude SNPs from consideration as tags, ensuring that all chosen tag SNPs have validation scores above 0.6. Tagger does not provide a list of alternative tag SNPs (or combinations of tag SNPs) from which one could choose those with the highest validation

score. It is possible that another selection of tag SNPs would perform equally well in capturing allelic variation in the gene at the given r² threshold with uniformly better validation scores. The only way one could check this is to further exclude SNPs with lower validation scores, rerun the program, and determine whether the next selection of tag SNPs all have higher validation scores while still satisfying the r² threshold. Clearly this has the potential to affect the combination of tag SNPs selected and the groups of SNPs represented by them.

Results

A total of 35 genes (see Table 1) were chosen for examination and genotyping in subjects from the HealthIron study. This selection was based on prior biological evidence for involvement in iron metabolism. Some of these genes have strong historical evidence of involvement in

Table 1: HealthIron SNPs selected for typing by gene, incomplete cover means not all tag SNPs were included.

No. SNPs		HUGO	Full or other name(s)	Chr
16	*	ACO1	IRP1; IRE-binding protein; Aconitase 1	9p21.2
1		CALR	Calreticulin	19p13.13
14		CD163	Hemoglobin scavenger receptor	12p13.31
25		CP	Ceruloplasmin	3q25.1
29		CUBN	Cubulin (incomplete cover)	10p13
27	*	CYBRD1	Dcytb	2q31.1
6		DHCR7	Smith-Lemli-Opitz syndrome	11q13.4
3		EXOC6	SEC15 (incomplete cover)	10q23.33
16		FLVCR	Feline leukemia virus subgroup C receptor	1q32.3
7	*	FTH1	H-ferritin	11q12.3
1		FTL	L-Ferritin	19q13.33
7		FXN	Frataxin	9q21.11
2		GAST	Gastrin	17q21.2
1		GSTP1	GSTP1 I105V	11q13.2
5	*	HAMP	Hepcidin	19q13.12
5	*	HCP1	MGC9564, heme carrier protein 1	17q11.2
18	*	HEPH	Hephaestin	Xq12
9		HEPHL1	Hephaestin-like 1; Eleutherin	11q21
7		HFE	Hemochromatosis	6p22.2
4	*	HFE2	Hemojuvelin	1q25
10		HMOX1	HO-1Hemoxygenase 1	22q12.3
2		HMOX2	HO-2 Hemoxygenase 2	16p13.3
2		HP	Haptoglobin	16q22.2
6		HPX	Hemopexin	11p15.4
18	*	IREB2	IRP2 IRE-binding protein 2	15q24.1
1	*	PGRMC1	Progesterone receptor membrane component 1	Xq24
3	*	PGRMC2	Progesterone receptor membrane component 2	4q28.2
13		SLC11A2	Divalent metal-ion transporter 1; DCT1	12q13.12
20		SLC25A37	Frascati	8p21.2
12		SLC40A1	Ferroportin; IREG1; MTP1	2q32.2
11		STEAP3	nm1058 (incomplete cover)	2q14.2
45	*	TF	Transferrin	3q22.1
6	*	TFR2	Transferrin receptor 2	7q22.1
30	*	TFRC	Transferrin receptor 1	3q29
2		TNF	TNFalpha	6p21.33
384				

* indicates genes where NHBLI RS&G Caucasian data was used.

cellular iron uptake and storage (*TF, TFRC, FTH1, FTL*) whereas others were chosen based on more recent evidence (*SLC40A1, SLC11A2, CP, TFR2, CYBRD1, IREB2, HFE2, HEPH, HEPHL1, HAMP, HCP1, DHCR7, HP, HMOX1*). Several genes (*CUBN, STEAP3* and *EXOC6*) were not completely covered i.e. not all the required tag SNPs were included as we had reached our total of 384 SNPs.

A substantial percentage of SNPs (41%) we submitted for validation had scores below 0.6, and were excluded from being selected as tag SNPs. This meant coverage of SNPs was not complete, although in regions with high LD there was usually an alternative SNP to tag those with low validation scores. A second round of genotyping on a different platform will be performed for the HealthIron Study to attempt to capture these low validation uncaptured SNPs. Figure 1 shows that the relationship between validation scores for SNPs and "Gentrain" score (a measure of SNP performance automatically calculated by the Illumina BeadStudio software) is not strong. Half of the ten "unscorable" SNPs were Golden Gate validated (previously successful), i.e. given scores of 1.1; overall 35% of our selected SNPs had scores of 1.1. This suggests that validation/design scores above 0.6 do not predict genotyping performance, and that maximising average validation score may not have a large effect on SNP success rate.

Table 2 displays, for each gene and for data on Caucasian individuals only from each data source, a comparison of the number of tag SNPs selected by LDselect versus Tagger in Haploview where the r^2 threshold for capturing SNPs was set at 80%. The column labelled "difference" shows

the actual increase in efficiency by Tagger due to its ability to replace some tag SNPs with multi-marker combinations of other tag SNPs [11]. The ratio of the number of tag SNPs to total number of variable SNPs (the last column of Table 2) is a reflection of the LD variation in iron homeostasis genes. Table 3 examines the effect of varying the r^2 cutoff on the number of tagSNP for the six largest genes that were resequenced. The result from Tagger can vary betweens runs with identical settings so the lowest result from 10 runs is reported.

Resequencing data versus HapMap
Table 4 compares the number of tag SNPs (using LDselect) selected from NHLBI resequencing data and HapMap (both Caucasian samples). Using a minor allele frequency (MAF) $\geq$ 3%, there were two genes, *HFE2* and *FTH1*, for which there were no variable HapMap SNPs in the CEU population. Averaged across all the genes (in Table 4 the sum of NHLBI resequencing tag SNPs column divided by sum of HapMap tag SNPs column) there were 2.8 times as many tag SNPs required for resequencing cover than HapMap phase 1 due to the larger number of SNPs, especially of low frequency.

Analysis of genotyping data for rs2239641 showed four data clusters rather than the usual three (for minor allele homozygotes, heterozygotes and major allele homozygotes). Using the resequencing data we were able to show that this inconsistency was due to a nearby (14 bp) untyped SNP causing two heterozygote clusters. For another pair of SNPs (TNFalpha rs1800630 and rs1800610) there were no compound heterozygotes. This may be due to an untyped SNP (rs1799724) 6 bp from
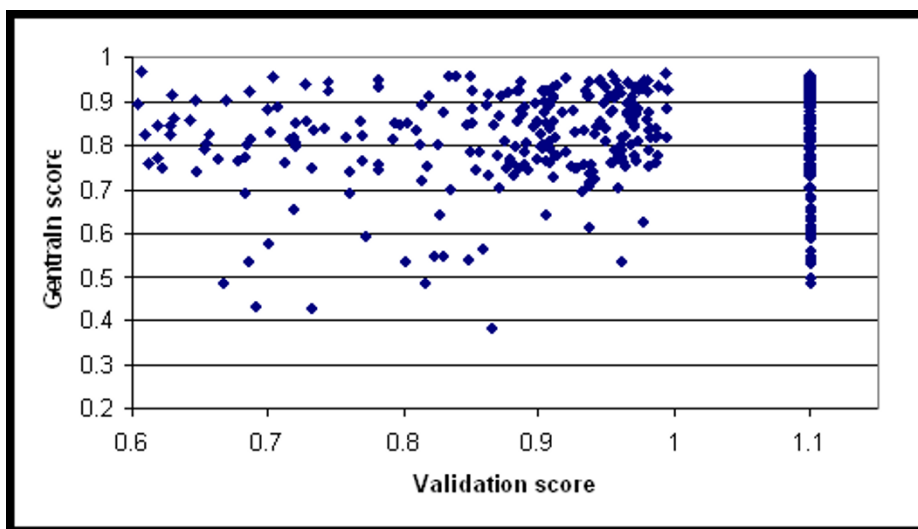
Figure 1
Scatter plot of SNP validation scores with Gentrain scores. Previously successful SNPs are given a score of 1.1, design scores between 0 and 1 are calculated by a proprietary algorithm based on the surrounding 200 bp.

Table 2: Comparison of number of tag SNPs generated by LDselect and Tagger in Haploview using Caucasian datasets (HAPMAP, NHLBI, Seattle SNPs). A "test" in Tagger refers to a combination of one to three tag SNPs used to capture one or more SNPs. The number of tests is the number of tag SNPs combinations required to capture all SNPs with a minimum value of r^2 of 80%. HapMap Phase 1 data was all that was available at the time of SNP selection.

		No. SNPs	Tagger	LDselect	Difference	Tags/Total Tagger	Tags/Total LDselect
Gene	Data Source	MAF > 3%	Tags/Tests	Tags			
ACO1	HapMap	27	13/15	15	2	0.48	0.56
CP	HapMap	11	8/8	8	0	0.73	0.73
CYBRD1	HapMap	10	5/5	5	0	0.50	0.50
	NHLBI RS&G	62	14/14	16	2	0.23	0.26
FLVCR	HapMap	17	6/8	8	2	0.35	0.47
	NHLBI RS&G	99	16/19	17	1	0.16	0.17
FTH1	NHLBI RS&G	13	5/5	5	0	0.38	0.38
HAMP	NHLBI RS&G	7	4/4	4	0	0.57	0.57
HCP1	NHLBI RS&G	6	3/3	3	0	0.50	0.50
HEPH	HapMap	40	6/6	6	0	0.15	0.15
IREB2	HapMap	12	4/5	5	1	0.33	0.42
SLC11A2	HapMap	13	4/4	4	0	0.31	0.31
SLC40A1	HapMap	6	3/4	4	1	0.50	0.67
TF	HapMap	11	7/7	7	0	0.64	0.64
	NHLBI RS&G	43	17/20	20	3	0.40	0.47
	SeattleSNPs	101	29/31	31	2	0.29	0.31
TFR2	NHLBI RS&G	6	4/4	4	0	0.67	0.67
TFRC	HapMap	12	6/6	6	0	0.50	0.50
	NHLBI RS&G	117	33/38	36	3	0.28	0.31

rs1800630 causing the assay to fail when it is present. Illumina specifies a distance of 60 bp to reject two adjacent SNPs being included in the same OPA (oligo pool all). Resequencing can thus provide an additional exclusion criterion to avoid choosing as tagSNPs any SNP with another common SNP within 60 bp.

Transferrin as an example of the effect of resequencing coverage for tagSNP selection

Here we present an empirical example of the effect of resequencing coverage on the selection of tag SNPs. Figure 2 shows the coverage of the transferrin gene (*TF*) by the four data sources. HealthIron resequencing data used only the exons and small amounts of the surrounding introns (at least 30 bp). The NHLBI RS&G data had a wider coverage (green) and SeattleSNPs coverage was nearly complete (black), with HapMap CEU SNPs in brown.

Table 3: Effect of varying r^2 on number of tagSNPs selected using Tagger and MAF≥3% for six genes using NHLBI resequencing data. First number is tagSNPs using 2 and 3 multi-marker SNP tagging, second number is number of tests (the number of unique combinations of tag SNPs required to represent all SNPs).

r^2	TF	CYBRD1	ACO1	FLVCR	IREB2	TFRC
0.50	10/10	9/9	19/21	8/8	10/11	22/22
0.55	10/10	10/10	20/21	9/9	11/11	22/22
0.60	11/11	12/14	21/23	10/10	11/11	25/25
0.65	12/13	14/16	22/23	10/11	11/11	26/26
0.70	14/16	15/15	22/23	11/12	11/11	27/28
0.75	16/18	15/15	23/25	13/14	12/12	29/29
0.80	17/19	15/15	23/26	15/17	12/12	33/36
0.85	19/20	16/16	26/27	16/18	12/12	37/38
0.90	19/21	18/21	27/32	17/18	12/12	42/43
0.95	22/23	21/25	31/35	19/20	13/13	46/51
1.00	23/36	21/25	34/37	22/27	14/15	50/52

Table 4: Number of tag SNPs from NHLBI resequencing versus HapMap data (Phase 1) for Europeans using MAF ≤ 3% and r^2 0.8.

Gene	NHLBI RS&G	HapMap Phase 1
ACO1	26	15
CYBRD1	16	5
FLVCR	17	8
FTH1	5	0
HAMP	4	1
HCP1	3	1
HEPH	13	6
HFE2	5	0
IREB2	11	5
PGRMC1	5	1
PGRMC2	2	2
TF	20	7
TFR2	4	2
TFRC	36	6
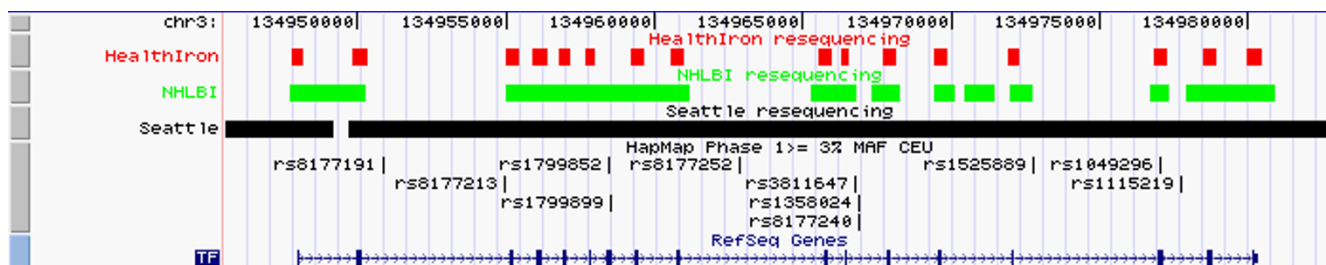
Figure 2
Regions sequenced in three resequencing Caucasian data sets: (i) HealthIron in red; (ii) NHLBI RS&G in green; (iii) SeattleSNPs in black. The HapMap Phase 1 Caucasian (European) SNPs with MAF ≥ 3% rs numbers are shown. The *TF* gene appears in blue with the exons shown as bars. The arrows indicated the direction of transcription.

Table 5 compares the number of SNPs with MAF ≥ 3% identified in data from Caucasians, using the four data sources (HealthIron, NHLBI RS&G, SeattleSNPS, and HapMap). The 7 HapMap tag SNPs, chosen from the 11 HapMap SNPs with MAF ≥ 3% using Tagger, capture just 45 out of the 101 Seattle SNPs (45%) using an r^2 threshold of 80% (so only 45% of SNPs have an r^2 of 0.80 or more with at least one tag SNP). Increasing the minimum MAF to 5% increases the capture of Seattle SNPs using HapMap tag SNPs to 48/89 (55%). Decreasing the r^2 threshold to 0.50 with minimum MAF still 3% only improved capture of the Seattle SNPs to 63/101 (62%). Approximately 55% of variant SNPs for transferrin in the SeattleSNPs database were not captured well using only HapMap phase 1 tag SNPs. In comparison HapMap Mar 2006 has 38 SNPs within the TF gene (MAF ≥ 3% in Caucasians), with 17 tag SNPs (pairwise using Tagger). Unfortunately *TF* was the only gene for which we could make this comparison, since it requires all the HapMap SNPs to be within the regions for which resequencing data are available. Figure 3 shows the minor allele frequency of the captured and uncaptured SNPs, showing that there was an even frequency distribution of SNPs not captured, not just low frequency.

Figure 4 shows the additional detail that is available with increasing coverage of the SNP data in the same genomic region, from the coarse grain of the HapMap SNPs through to the fine grain of SeattleSNPs which approaches complete resequencing. While the major feature of a large block of LD on the righthand side of the display is apparent at all levels, there is much more detail with resequencing data and additional blocks of LD are revealed as more SNPs are added.

Linkage disequilibrium patterns across five populations
Table 6 lists the number of tag SNP and total number of SNPs with MAF ≥ 3% for 14 genes resequenced by NHLBI across five population samples. The ratio of tag to total SNPs is low for genes within which there is substantial LD (e.g. *FLVCR*). The number of tag SNPs required is higher for African populations due to both the higher total number of SNPs and lower LD for most genes (as shown by higher ratios of tag SNPs to total SNPs). The final column uses multiPopTagSelect [12] to choose a minimal union of population-specific tag SNPs to capture SNP variation across all five populations. Figure 5 shows the pattern of linkage disequilibrium across five populations for the six genes which had more than 40 SNPs in each population. There is substantial variation in LD patterns both across genes and across populations which likely represents admixture variability throughout the genome.

We found the average capture rate of SNPs with MAF ≥ 3% across 10 genes (*TF, TFR2, CYBRD1, FTH1, HAMP, HFE2, HCP1, IREB2, FLVCR* and *ACO1*) in the Hispanic population using European tags was 82.5% (using NHLBI data i.e. 47 European descent, 48 Hispanic descent), and for

Table 5: The number of SNPs identified for the transferrin gene from the four different data sources with different coverage. Figure 1 shows the coverage of each data set, N is the number of Caucasian individuals sequenced/typed.

	Resequencing			HapMap	HapMap
	HealthIron N = 188	NHLBI RS&G N = 47	SeattleSNPs N = 23	N = 30 trios Phase 1	N = 30 trios Mar 2006
Total No. SNPs	31	59	128	12	81
No. SNPs ≥ 3% MAF	17	43	101	11	38
Tag SNPs with Tagger	13	20	31	7	17
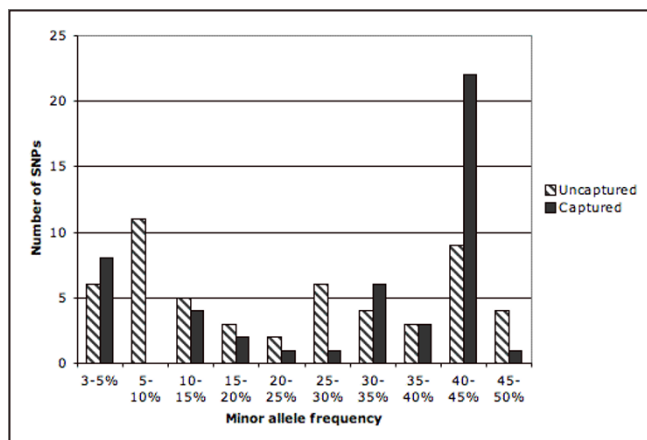
Figure 3
Frequency distribution of captured and uncaptured SNPs from Seattle resequencing of *TF* using HapMap tagSNPS. The large number of captured SNPs in the 40–45% range represents the strong block of LD which is captured by a single tagSNP.

the African American sample using Yoruban tags 78.1% (data not shown). However, there was one gene (*FTH1*) whose capture rate was < 50% for both comparisons. The capture rate of Yoruban SNPs using European tag SNPs was poor, an average of 35% over 10 genes, range 14–60% (data not shown).

Coverage of stock genotyping arrays
To compare our findings with the coverage that would be achieved with the "standard" Affymetrix and Illumina genotyping arrays, the number of SNPs appearing across various genotyping arrays within the regions for which we did partial resequencing are displayed in Table 7. While this is not a direct comparison of true coverage, which would require complete resequencing information, arrays with less than the number of tagSNPs cannot meet our coverage criteria of $r^2 \geq 0.8$ for SNP with MAF $\geq 3\%$. TFRC needed a higher number of tagSNPs than the number of SNPs present even on the latest array versions. HCP1 (previously known as MGC9564) has no SNPs on Affymetrix arrays.

Discussion
We have reported on the challenges we faced in selecting SNPs from candidate genes of iron metabolism for genotyping in an association study. At the time we were required to make the SNP selection for the HealthIron study, resequencing data from the same participants was available for only 6 genes, and the coverage based on this local sample included only the exons and immediate surrounding genomic regions. We turned instead to several public available databases, including the International
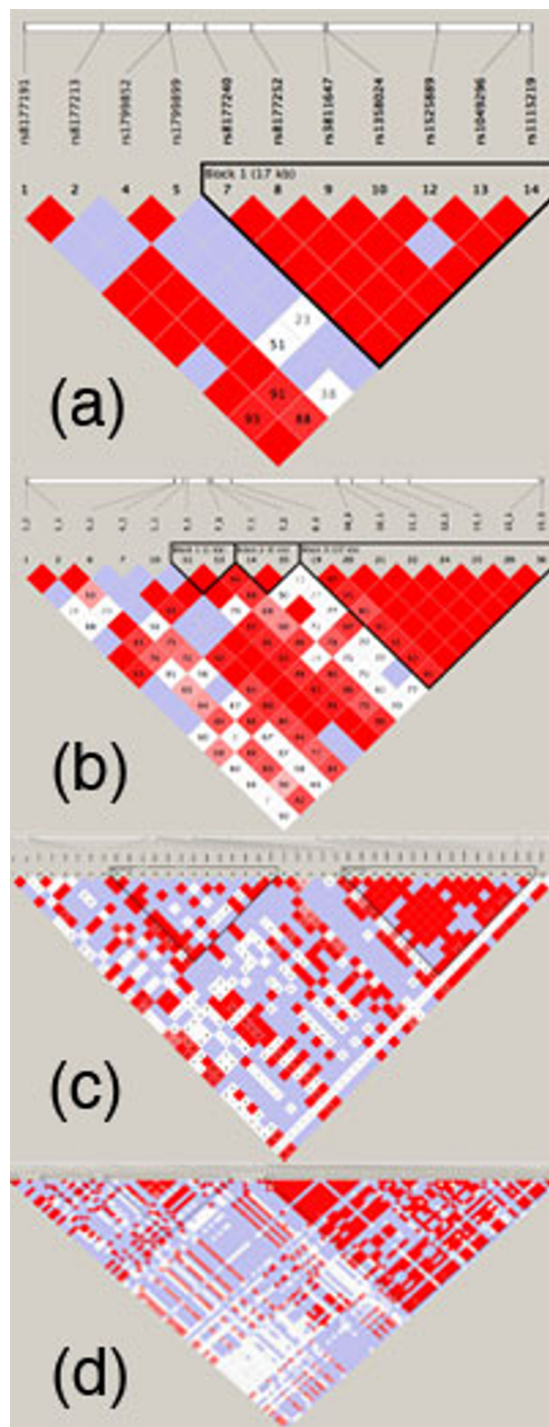
Figure 4
A graphical representation of linkage disequilibrium patterns for the transferrin gene from SNP data on Caucasian populations: (a) HapMap (11 SNPs with MAF $\geq$ 3%); (b) HealthIron (12 SNPs); (c) NHLBI RS&G (43 SNPs); (d) SeattleSNPs (101 SNPs). These LD displays were generated using the default settings in HaploView.

Table 6: Number of tag SNPs/total SNPs with MAF ≥ 3% (ratio) in each of five populations using LDSelect with NHLBI resequencing data. The last column shows number of tag SNPs and total SNPs combined across the five populations using multiPopTagSelect.

Gene	European N = 47	Hipanic N = 48	Chinese N = 47	African American N = 24	Yoruban N = 24	Multiple
ACO1	26/58 (0.45)	29/56 (0.52)	28/57 (0.49)	69/111 (0.62)	65/110 (0.59)	97/145 (0.67)
CYBRD1	15/62 (0.24)	20/58 (0.34)	17/62 (0.27)	28/57 (0.49)	34/65 (0.52)	46/85 (0.54)
FLVCR	17/99 (0.17)	11/75 (0.15)	8/83 (0.10)	18/82 (0.22)	18/91 (0.20)	27/114 (0.24)
FTH1	5/13 (0.38)	10/17 (0.59)	4/7 (0.57)	11/19 (0.58)	8/14 (0.57)	17/27 (0.63)
HAMP	4/7 (0.57)	4/7 (0.57)	7/10 (0.70)	8/10 (0.80)	9/11 (0.82)	11/15 (0.73)
HCP1	3/6 (0.50)	5/9 (0.56)	4/9 (0.44)	11/16 (0.69)	11/15 (0.73)	17/22 (0.77)
HEPH	13/38 (0.34)	10/36 (0.28)	9/10 (0.90)	16/58 (0.28)	21/53 (0.40)	49/95 (0.52)
HFE2	5/7 (0.71)	5/8 (0.63)	4/7 (0.57)	8/10 (0.80)	5/7 (0.71)	10/13 (0.77)
IREB2	11/52 (0.21)	11/47 (0.23)	15/63 (0.24)	21/94 (0.22)	18/95 (0.19)	42/134 (0.31)
PGRM1	5/10 (0.50)	2/2 (1.00)	0/0	8/11 (0.73)	10/13 (0.77)	15/19 (0.79)
PGRM2	2/4 (0.50)	4/8 (0.50)	7/9 (0.78)	10/13 (0.77)	11/17 (0.65)	16/22 (0.73)
TF	20/44 (0.45)	24/50 (0.48)	17/49 (0.35)	34/61 (0.56)	42/61 (0.69)	61/87 (0.70)
TFR2	4/6 (0.67)	3/5 (0.60)	5/7 (0.71)	5/5 (1.00)	6/7 (0.86)	8/9 (0.89)
TFRC	36/119 (0.30)	34/115 (0.30)	22/93 (0.24)	43/135 (0.32)	44/110 (0.40)	87/187 (0.47)
Av No. tags	11.9	12.3	10.5	20.7	21.6	35.9

HapMap and Seattle SNPs projects, and the NHLBI RS&G data analysed for the HEIRS ancillary study on iron deficiency. Data from the NHLBI RS&G project identified SNPs for 14 genes (asterisked in Table 1). There were 32 novel SNPs not in dbSNP (so without *rs* numbers) included in the HealthIron tag SNP list.

The benefits of resequencing for genetic data include detection of novel SNPs, a more detailed coverage of the candidate genes, as indicated by the higher number of tag

SNPs generated from resequencing data in comparison to the HapMap (2.8 times over 14 genes), and knowledge of the precise pattern of LD in the population of interest. We included some genes in our study that did not have any variable SNPs in the HapMap phase 1 database.

The transferability of tag SNPs has received considerable attention recently, with reports of good performance of European (CEU) tag SNPs in populations from the United Kingdom [13], Finland, Estonia [14]and the Pacific Rim
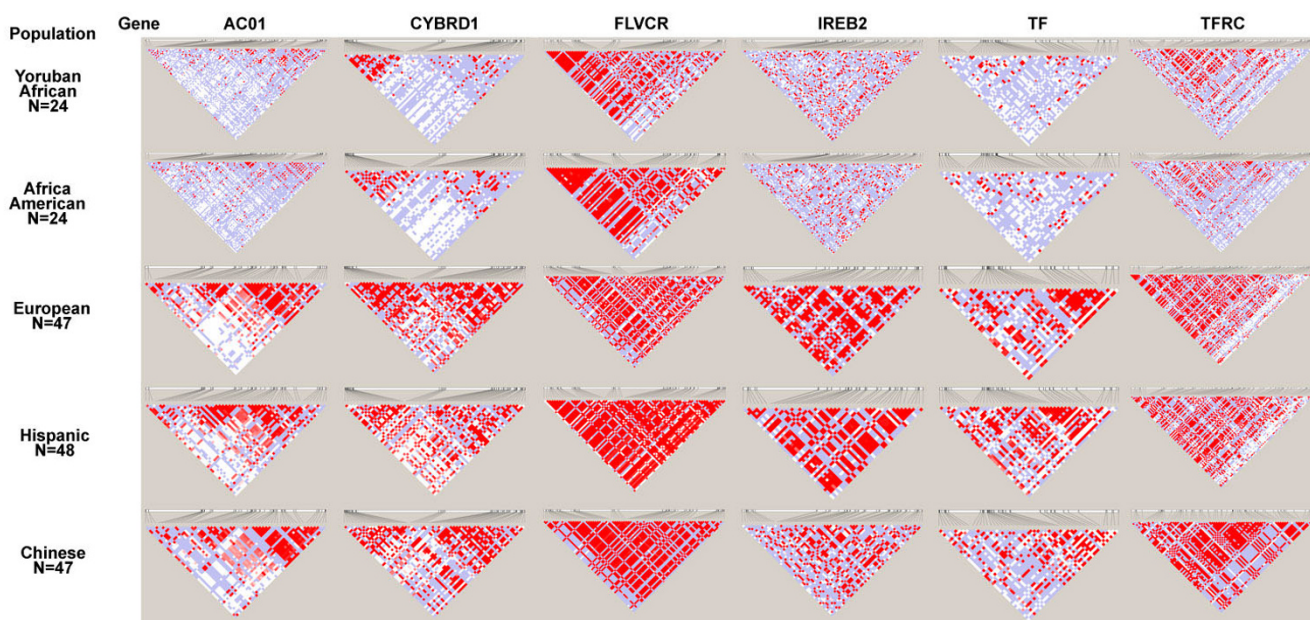
Figure 5
Pattern of linkage disequilibrium across six genes and five population samples using Haploview default settings (with blocks removed).

Table 7: Number of SNPs on different genotyping arrays within each gene region of NHLBI resequencing. Note that in all cases the entire region was not resequenced and so the first two columns are conservative and refer only to the sample of 47 people of European descent.

Gene	NHLBI reseq total SNPs	NHLBI tagSNPs MAF≥3% r2 0.8	Region included in basepairs	Affy 100K	Affy 500K	Affy GW5	Affy GW6	Illumina Humhap 300	Illumina Humhap 550	Illumina Human 1M
ACO1	276	26	69937	1	11	12	24	14	26	41
CYBRD1	139	16	39535	0	3	3	8	7	12	19
FLVCR	185	17	40442	4	6	6	15	11	12	18
HAMP	43	4	6314	0	2	2	3	1	1	4
HCP1	65	3	11169	0	0	0	0	1	2	7
HEPH	177	13	101478	1	11	10	32	5	7	21
HFE2	38	5	7437	0	1	1	2	0	1	5
IREB2	241	11	64450	1	15	11	21	3	7	24
PGRMC1	33	5	12051	0	0	0	1	1	1	7
PGRMC2	55	2	21640	0	1	1	4	2	3	9
TF	158	20	33189	0	23	18	37	8	11	23
TFR2	56	4	21649	0	0	0	1	3	3	13
TFRC	314	36	36402	2	4	4	10	7	8	14

(in particular Japanese, Chinese, Hawaiian and Latino)[15]. We confirmed the finding of de Bakker et al. [11] that European tag SNPs performed poorly in the African population, capturing only an average of 35% of SNP variation. Our results are also consistent with those of Tantoso et al. [16], who found that the SNPs in the Hap-Map capture only about 55% of untyped variants. For the transferrin gene, using the SeattleSNPs database we found 45% of SNPs with MAF ≥ 3% were captured by HapMap tag SNPs.

Although the cross-population tagging rate for the Hispanic population using European tags (83%) and African American using Yoruban tags (78%) was high for nine of the ten genes, there was one gene (*FTH1*) for which the capture rate was less than 50%. For candidate gene/region studies where there is strong a priori evidence of association, an exhaustive SNP search is desirable and hence using tag SNPs selected from a different population may result in a failure to genotype any markers in strong LD with a causative variant. This variability is likely to represent admixture variability throughout the genome. Some of these population differences may reflect different selection history among populations due to diet or disease prevalence that may be relevant in a genetic association study.

A sensible approach to overcoming the problem in of selecting tag SNPs in multi-ethnic cohort studies is to use an optimal union of population-specific tag SNPs as implemented in the program multiPopTagSelect [12]. Although this resulted in a three fold increase in the number of tag SNPs required to capture all five populations in the NHLBI RS&G database compared to selected

tag SNPs in the European sub-population alone, it was a still a substantial reduction compared to the additional genotyping that would have been required if selection had been performed independently within each population. The algorithm proposed by Howie et al. (2006) leverages the existing results of LDselect within separate population samples, and can easily accommodate tag SNPs ranked by a performance criterion such as the validation scores with which we were provided for use with the Illumina genotyping platform. In contrast, it was time-consuming to determine "manually" the set of tag SNPs with the highest possible combined validation score using Tagger, since it makes no use of these scores with the exception that SNPs with validation scores below a threshold can be excluded from consideration. In hindsight it would have been quicker to use LDselect alone for this particular collection of candidate genes, as the difference in tag SNPs selected was small relative to the amount of time spent rerunning Tagger. The previously reported increase in efficiency of Tagger using two and three tag SNPs combinations [11] also potentially requires more complex analysis than single SNP associations. Tagger should be rerun several times as the number of tagSNPs required may vary between runs with identical settings.

Conclusion
When selecting SNPs for genotyping in association studies, the candidate gene approach is distinct from the whole genome scanning approach which only examines common SNPs. For candidate genes it is preferable to augment HapMap data with sequence data. This has the advantage of aiding in the discovery and coverage of SNPs with frequencies in the range of 1% to 5% which are unlikely to appear in the HapMap database. Remaining

challenges in SNP selection within candidate genes include developing methods for combining multiple sources of information and incorporating redundancy to overcome platform limitations.

Competing interests
We have no interests to declare but note that SMF is the director of the Australian Genome Research Facility.

Authors' contributions
DMG, CDV, GJA, CEM, MB, SF, and KJA were involved in study concept, design and reviewed the manuscript. CCC and LCG analysed the genetic data and with CEM drafted the manuscript. GJA and CDV created the list of candidate genes. All authors read and approved the final manuscript.

References
1. Thorisson GA, Smith AV, Krishnan L, Stein LD: **The International HapMap Project Web site.** *Genome Res* 2005, **15(11):**1592-1593.
2. Consortium TIHM: **The International HapMap Project.** *Nature* 2003, **426(6968):**789-796.
3. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R Jr., Ellis MC, Fullan A, Hinton LM, Jones NL, Kimmel BE, Kronmal GS, Lauer P, Lee VK, Loeb DB, Mapa FA, McClelland E, Meyer NC, Mintier GA, Moeller N, Moore T, Morikang E, Prass CE, Quintana L, Starnes SM, Schatzman RC, Brunke KJ, Drayna DT, Risch NJ, Bacon BR, Wolff RK: **A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis.** *Nat Genet* 1996, **13(4):**399-408.
4. Giles GG, English DR: **The Melbourne Collaborative Cohort Study.** *IARC Sci Publ* 2002, **156:**69-70.
5. McLaren CE, Barton JC, Adams PC, Harris EL, Acton RT, Press N, Reboussin DM, McLaren GD, Sholinsky P, Walker AP, Gordeuk VR, Leiendecker-Foster C, Dawkins FW, Eckfeldt JH, Mellen BG, Speechley M, Thomson E: **Hemochromatosis and Iron Overload Screening (HEIRS) study design for an evaluation of 100,000 primary care-based adults.** *Am J Med Sci* 2003, **325(2):**53-62.
6. SeattleSNPs: **NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA.** [http://pga.gs.washington.edu].
7. **International HapMap Project Homepage** [http://www.hapmap.org/index.html.en]
8. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: **Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.** *Am J Hum Genet* 2004, **74(1):**106-120.
9. Nickerson DA, Rieder MJ, Carlson CS, Yi Q: **ldSelect.** [http://droog.gs.washington.edu/ldSelect.html].
10. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21(2):**263-265.
11. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: **Efficiency and power in genetic association studies.** *Nat Genet* 2005, **37(11):**1217-1223.
12. Howie BN, Carlson CS, Rieder MJ, Nickerson DA: **Efficient selection of tagging single-nucleotide polymorphisms in multiple populations.** *Hum Genet* 2006, **120(1):**58-68.
13. Ke X, Durrant C, Morris AP, Hunt S, Bentley DR, Deloukas P, Cardon LR: **Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples.** *Hum Mol Genet* 2004, **13(21):**2557-2565.
14. Montpetit A, Nelis M, Laflamme P, Magi R, Ke X, Remm M, Cardon L, Hudson TJ, Metspalu A: **An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population.** *PLoS Genet* 2006, **2(3):**e27.
15. de Bakker PI, Graham RR, Altshuler D, Henderson BE, Haiman CA: **Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations.** *Pac Symp Biocomput* 2006:478-486.
16. Tantoso E, Yang Y, Li KB: **How well do HapMap SNPs capture the untyped SNPs?** *BMC Genomics* 2006, **7:**238.

Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2350/9/18/prepub