BMC Genomics

BioMed Central

Research article

Open Access

Genome annotation of a 1.5 Mb region of human chromosome 6q23 encompassing a quantitative trait locus for fetal hemoglobin expression in adults

James Close[1,2], Laurence Game[1,3], Barnaby Clark[1], Jean Bergounioux[1,4], Ageliki Gerovassili[1] and Swee Lay Thein*[1]

Address: [1]Department of Haematological Medicine, GKT School of Medicine, King's Denmark Hill Campus, Bessemer Road, London, SE5 9PJ, UK, [2]SANE POWIC, Warneford Hospital, Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, UK, [3]CSC-IC Microarray Centre, 2nd floor, L-block, Room 221, Imperial College Faculty of Medicine, Hammersmith Hospital Campus, Du Cane Road, London, W12 0NN, UK and [4]Unité de soins intensif pédiatrique, Hôpital Universitaire Krémlin Bicêtre, 63 av. Gabriel Péri, 94270 Le Krémlin Bicêtre, France

Email: James Close - james.close@psych.ox.ac.uk; Laurence Game - laurence.game@imperial.ac.uk; Barnaby Clark - barnaby.clark@kingsch.nhs.uk; Jean Bergounioux - bergouniouxj@magic.fr; Ageliki Gerovassili - ageliki.2.gerovassili@kcl.ac.uk; Swee Lay Thein* - sl.thein@kcl.ac.uk

* Corresponding author

Published: 31 May 2004

BMC Genomics 2004, **5**:33

Received: 13 February 2004
Accepted: 31 May 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/33

Abstract

Background: Heterocellular hereditary persistence of fetal hemoglobin (HPFH) is a common multifactorial trait characterized by a modest increase of fetal hemoglobin levels in adults. We previously localized a Quantitative Trait Locus for HPFH in an extensive Asian-Indian kindred to chromosome 6q23. As part of the strategy of positional cloning and a means towards identification of the specific genetic alteration in this family, a thorough annotation of the candidate interval based on a strategy of *in silico* / wet biology approach with comparative genomics was conducted.

Results: The ~1.5 Mb candidate region was shown to contain five protein-coding genes. We discovered a very large uncharacterized gene containing WD40 and SH3 domains (*AHI1*), and extended the annotation of four previously characterized genes (*MYB*, *ALDH8A1*, *HBS1L* and *PDE7B*). We also identified several genes that do not appear to be protein coding, and generated 17 kb of novel transcript sequence data from re-sequencing 97 EST clones.

Conclusion: Detailed and thorough annotation of this 1.5 Mb interval in 6q confirms a high level of aberrant transcripts in testicular tissue. The candidate interval was shown to exhibit an extraordinary level of alternate splicing – 19 transcripts were identified for the 5 protein coding genes, but it appears that a significant portion (14/19) of these alternate transcripts did not have an open reading frame, hence their functional role is questionable. These transcripts may result from aberrant rather than regulated splicing.

Background

The hemoglobinopathies represent the most common category of clinically significant inherited disorders, causing a huge burden on global health [1]. Despite the apparently simple Mendelian inheritance of α- and β-thalassemia and sickle cell disease (SCD), a significant variation in clinical severity is observed. It is now evident that the genetic background of affected individuals imparts a substantial portion of the variation in clinical phenotype [2]. In particular, a number of studies have shown that an increased fetal hemoglobin (HbF; $\alpha^2\gamma^2$) response in conjunction with either sickle cell disease or β-thalassemia has an ameliorating effect on the clinical disease [3,4]. This has prompted intensive investigations on γ-globin gene regulation, the outcome of which may provide insights for the therapeutic augmentation of HbF as treatment for the β-hemoglobinopathies.

All normal adults continue to synthesize small quantities of HbF; the residual amounts of HbF are restricted to a subset of erythrocytes termed F cells [5]. Surveys have shown that the distribution of HbF in adults is continuous and varies considerably (>20 fold) between individuals [6]. The heritability of FC levels has been estimated to be 0.89 in the European population [7]. Whilst environmental influences – including age [8] and sex [8,9] – account for a proportion of this variation, family studies have demonstrated a strong genetic component of this variation [6]. These genetic influences include DNA sequence variants *in-cis* to the β-globin complex such as the C→T single base substitution at position -158 in the promoter of the $^{G}\gamma$-globin gene (referred to as the *Xmn1*-$^{G}\gamma$ site [7,10]). The major proportion of the variance in HbF, however, is due to *trans*-acting factors [11]. About 10–15% of individuals have levels of HbF > 1% and up to 5% of total hemoglobin. Such individuals with modest increases in HbF are considered to have heterocellular HPFH in which there is uneven distribution of HbF among the erythrocytes (hence heterocellular) [12]. It is likely that several quantitative Trail Loci (QTLs) contribute to the HbF levels in heterocellular HPFH, unlike the rare pancellular HPFHs which are inherited in a Mendelian fashion as alleles of the β-globin complex, caused either by extensive deletions of the β-globin cluster or point mutations in the γ-globin promoter [12]. Heterozygotes for such pancellular HPFHs have a clearly defined phenotype of substantially increased HbF levels of 10–35% which is homogenously distributed among the red blood cells (hence 'pancellular').

Our group mapped a QTL for heterocellular HPFH in a very large Asian-Indian pedigree [13,14]. The kindred was initially identified through the proband who – despite being homozygous for β°-thalassemia, with a complete absence of adult hemoglobin (HbA; $\alpha_2\beta_2$) – had an extremely mild clinical phenotype which was the result of substantially raised HbF expression. The family was later characterized as a 210 member Asian-Indian kindred with heterocellular HPFH, β-thalassemia and α-thalassemia segregating through seven generations [15]. A genome-wide linkage search was performed on this family, leading to a significant linkage (Lod score = 12.4) between the chromosomal region 6q22.3-q24 and increased HbF expression [14,15]. In collaboration with the Sanger Centre, Cambridge, we assisted in the construction of a high-density BAC/PAC physical map covering the candidate interval [16]. The sequence for this candidate interval was subsequently produced by the Human Genome Mapping Project [17]. This paper describes a detailed annotation of the 1.5 Mb candidate interval, which was subsequently used to direct our mutation analysis studies in the Asian-Indian family.

Results

The annotated candidate interval is represented pictorially in Figure 1, detailing the approximate location of the five protein-coding genes and four RNA genes.

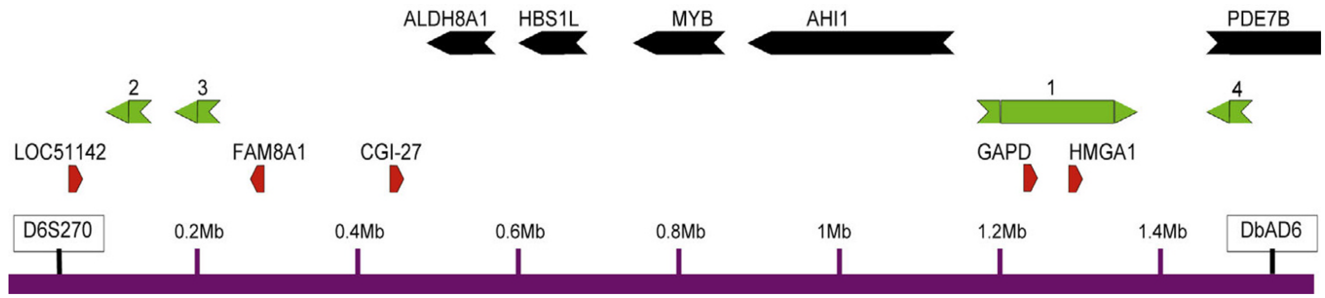

Figure 1
Annotation of the 1.5 Mb candidate interval. Protein coding genes are in black on top, with non-coding RNA genes indicated in green below. Pseudogenes are illustrated in red. Flanking markers are indicated on the scale (approx.). The direction of the arrows indicates transcriptional direction (5'→3') of the genes.

Characterization of the candidate interval

The candidate interval is localized to the cytogenetic "G-band" 6q23.2 on chromosome 6, comprising 1571770 bp (~1.5 Mb) of DNA defined by the markers D6S270 (**Z16636**) and DbAD6 (**AJ606363**). The extent of the candidate interval was defined by haplotype analysis that has been previously published [16], and that has been further refined to reduce the candidate interval. The telomeric boundary is now defined by the novel microsatellite marker DbAD6 (data not shown). This marker is within the 1st intron of the gene *PDE7B*, such that only the promoter and exon 1 of this gene remain within the candidate interval. The GC content of the candidate region is approximately 0.39 and the repeat content is approximately 44%, both of which are less than the genome average (0.41 and 50%, respectively). High GC and repeat content are positively associated with gene density [17], which suggests that the candidate interval is gene poor, an assumption which proved to be correct. The candidate interval was subjected to analysis for repetitive RNA encoding gene families. tRNAscan-SE [18] revealed no evidence for tRNA genes within the candidate interval, and no rRNA or micro RNA genes were identified by means of homology searches.

Genes

All genes were thoroughly annotated using a strategy which was based on manual inspection of public data sources to inform laboratory experiments. This combined a mixture of EST driven transcript identification, gene prediction and comparative genomics to identify transcripts. RT-PCR was used to confirm the existence of any putative transcripts deriving from these methods. Finally, where necessary, RACE (Rapid Amplification of cDNA Ends) was then used to obtain full-length cDNA for previously unidentified transcripts. This thorough annotation strategy was performed in an attempt to exhaustively clone and characterize all the transcripts within the candidate interval, including those that may be novel, expressed at a low-level or restricted to a specific tissue. The results of this analysis is described below, initially on a gene-by-gene basis, followed by the overall results for EST and comparative approaches.

MYB

MYB is a well-characterized nuclear transcription factor, expressed to some degree in most major hematopoietic lineages (for review see [19]). The genomic locus of *MYB* is comprised of 15 exons spread over approximately 48.5 Kb in the center of the candidate interval [20]. This locus expresses a 3.2 Kb mRNA, with the major human product of *MYB* being a 636AA, 80 kDa protein. Various putative splice variants of *MYB* have been reported in the literature [20-23], although other than the "exon 9A" splice variant [21,24,25], their existence is controversial. This exon 9A splice variant produces a larger 89 kDa protein which is the result of an addition of 363 bp between exons 9 and 10 (exon 9A); with this variant representing less than 10–20% of the total *MYB* protein in all cell types examined thus far. Within this work, the exon 9A splice variant is referred to as "Exon 9Aii", due to the confusion in the literature between exon 9A (exon 9Aii) and a *different* exon 9A splice variant annotated on the sequence **U22376** (referred to herein as exon 9Ai)

Further to the characterized exon 9Aii alternate transcript, several other *MYB* transcripts have been reported in the literature and public DNA databases. As a means to comprehensively annotating the gene, we tested the existence of all suggested putative splice variants of *MYB* by RT-PCR. These splice variants were suggested from a number of different resources and are detailed in table 1. We designed primers to test the existence of these transcripts by RT-PCR such that PCR products spanned an exon boundary, with one primer in a known (major transcript exon) and the other primer positioned in a putative alternate exon. PCR results for all transcripts that showed a positive result in a reasonable number of cycles (<35) are shown in Figure 2. All positive PCR products were sequenced across the splicing junctions and confirmed to represent the expected *MYB* alternate splice variant.

This strategy confirmed the existence of eight *MYB* splice variants including the primary transcript. Seven of these are listed in table 1 (Exon 8', Exon 8A, Exon 9Ai, Exon 9Aii, Exon 10A, Exon 13A, and Exon 14A), and the full sequences for each splice variant can be obtained from the listed accession numbers. Furthermore, sequencing of the PCR products revealed that Exon 8' (which uses an alternate splice donor in exon 8, producing a *MYB* exon 8 which is 9 bp smaller) and Exon 8A appear to be commonly expressed together, providing a ninth splice variant which contains both Exon 8' and Exon 8A (**AJ606321**). The arrangement of these splice variants is illustrated in Figure 3A.

Interestingly, all the variants involved alternate splicing events at the 3' end of the gene. This region correlates with the transactivation and negative regulatory domains of the gene, suggesting that different splice variants may interact with different protein partners. Conversely, the 5' end of the gene (relating to the DNA binding domain) appears completely invariant. It is possible that some of these splice variants represent low-level aberrant transcription and are of questionable biological significance. For example, only exon 9Aii and exon 8' splice variants contain a full protein coding region, whereas all other alternate splice variants introduce a premature stop codon.

Table 1: Putative alternate splice variants of *MYB* and RT-PCR results testing their existence.

	Evidence								
Putative Alternate exon	Literature	Existing sequence	EST Evidence	Ensembl transcript	Gene Prediction	Conserved region with mouse	RT-PCR primer sequences:	Positive result by RT-PCR (PCR conditions)	Sequence Accession Number
1A					✓		F: GCCCGAAGACCCCGGCACAG R:TGCAACTTCCCAGCGCCGCG		
1B						✓	F: GCCCGAAGACCCCGGCACAG R:CCAACAGTCAGAGGATTTTCAAAG		
1C						✓	F: GCCCGAAGACCCCGGCACAG R:CAATGCATCCAGCAATTACG		
1D						✓	F: GCCCGAAGACCCCGGCACAG R:GGCCACTTGTTAGTCAGAGTCTT		
1E					✓		F: GCCCGAAGACCCCGGCACAG R:TACACATTTGGCCACCTTCC		
2'	[22]	X52125					F: GCCCGAAGACCCCGGCACAG R:TGCAGAAATAAGAATGGGTAGACG		
5A					✓		F: GCCCTGCTGTGCCACATTCAAA R:CGGAGCCTGAGCAAAACCCATC		
8'	[22]	X52125	✓				F: AAATATAGTCAATGTCCCTCAGCC R:ATGTGTGGTTCTGTGTCTGCTGT	✓ 2.5 mM/60C	AJ606320
8-extended	[23]	M13666	✓				F: AAATATAGTCAATGTCCCTCAGCC R:ACAAAATAATAGAAAAAAACCAAAATG		
8A	[22]	U22376, X52125		✓			F: AAATATAGTCAATGTCCCTCAGCC R:GGGCATCACTTCTCTTTTATTGTCT	✓ 1.5 mM, 60C	AJ606317
8A-extended			✓			✓	F: AAATATAGTCAATGTCCCTCAGCC R:GCACATGCTAGTTCGACCAC		
9Ai		U22376					F: ACAGAACCACACATGCAGCTACC R:TGAAACAGCACTGAGAGAGAGATG	✓ 2.5 mM, 65C	AJ606318
9Aii	[21,22]	X17469, X52126, U22376		✓	✓	✓	F: AAATATAGTCAATGTCCCTCAGCC R:ACAGGGGAACGCTTGGGAGT	✓ 1.5 mM, 62C	AJ606319
9B						✓	F: ACAGAACCACACATGCAGCTACC R:CACAATTTGGTTCCCTCCTC		
9C						✓	F: ACAGAACCACACATGCAGCTACC R:TATTCCTTGGCAAAAACTGC		
9D						✓	F: ACAGAACCACACATGCAGCTACC R:GGTCATTGCTGAAAAACTTGC		
10A		U22376					F: ACAGAACCACACATGCAGCTACC R:CATCCCTTGGCTTCTAATCATATAA	✓ 2.5 mM, 55C	AJ606322
11'					✓		F: ACAGAACCACACATGCAGCTACC R:CCTTTTGATAGCTGGGGTTACAGT		
13A		U22376					F: AAACTTCTTCTGCTCACACCACTG R:CTCACGCCTGTAATCCTAGCAC	✓ 1.5 mM, 60C	AJ606323
14'					✓		F: AAACTTCTTCTGCTCACACCACTG R:AAATACCAATGTTGGAATAGTAAT		
14A						✓	F: AAACTTCTTCTGCTCACACCACTG R:TGGCTCAGGATTAATTTGGAA	✓ 2.5 mM, 50C	AJ606324
14B						✓	F: AAACTTCTTCTGCTCACACCACTG R:TGCAATATGTTCAATAATACCATGTG		
14C						✓	F: AAACTTCTTCTGCTCACACCACTG R:TCTCGCTTATCCTGTAATGTGC		

The putative alternate splice variant is indicated in the 1st column, and the subsequent columns indicate lines of evidence that suggested the existence of this splice variant. The literature references and accession numbers for previously reported splice variants are provided. Ensembl transcripts were observed on the v.9.30a.1 (2/12/2002) Ensembl release. Gene predictions were suggested by Genscan, Fgenesh++ or Twinscan, and conserved regions were classified as intronic sequence of >70% homology over 100 bp with the syntenic mouse region. The RT-PCR primer sequences used to test the existence of each alternate exon are listed, followed by a column indicating whether this splice variant was detectable by RT-PCR, and the associated PCR conditions. The final column provides the accession number of our sequences for the confirmed alternate splice variant.

HBS1L

HBS1L was originally identified during a comparative genome hybridization study searching for chromosomal imbalances in pancreatic adenocarcinoma [26]. Whilst *MYB* was identified as the most likely candidate oncogene in this study, *HBS1L* was also discovered among the co-amplified genes in the chromosomal region of 6q24. The product of the *HBS1L* gene was shown to encode a 648AA polypeptide, with a predicted molecular mass of around 75 kDa. Phylogenetic studies suggested that *HBS1L* may be associated with translating ribosomes and aid in the

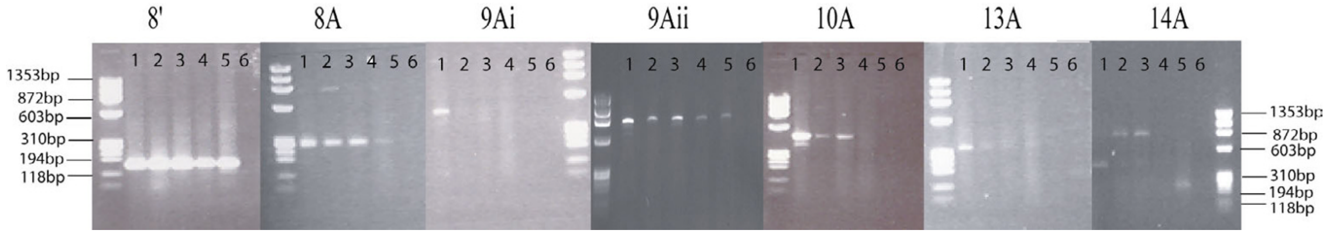

Figure 2

PCR results of confirmed *MYB* alternate splice variants. Alternate-exon specific primer pairs were tested against the same panel of cDNA tissues (L-R): 1) Erythroid, 2) Lymphoblastoid cell line, 3) Bone Marrow, 4) Fetal Liver, 5) Testes and 6) negative control. The final PCR (Exon 14A) exhibits aberrant products (sequence confirmed) in Lymphoblastoid cell line, Bone Marrow, and Testes, whereas the correct product (sequence confirmed) is observed in erythroid. The expected product sizes are 147 bp (8'), 260 bp (8A), 435 bp (9Ai), 738 bp (9Aii), 451 bp (10A), 324 bp (13A), and 450 bp, (14A). The size standard marker is øX174:HaeIII (Sigma), with some (but not all) of the band sizes indicated either side of the figure.

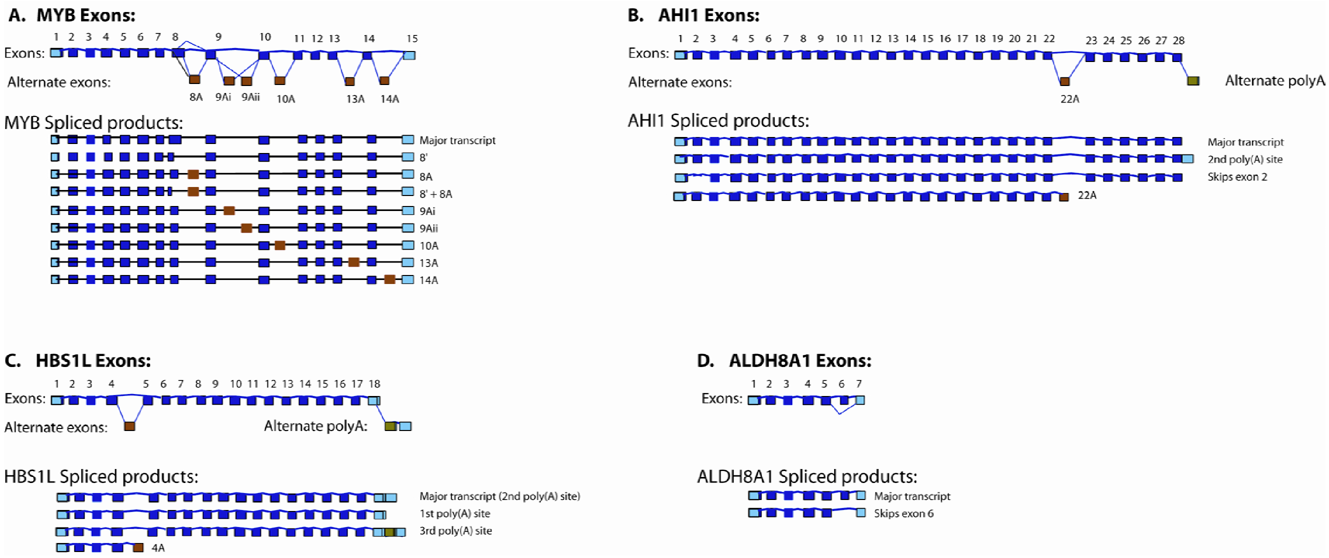

Figure 3

Alternate splicing of protein coding genes in candidate interval. (A) *MYB*, (B) *AHI1*, (C) *HBS1L* and (D) *ALDH8A1*. Internal exons are in dark blue, alternate exons in red and terminal exons in light blue. Each of the gene pictures displays the total pool of exons at the top, with the discovered splice variants displayed below, named according to the splice sites utilized.

passage of the nascent polypeptide through the ribosome channel. Alternatively, it may bring the amino-acyl-tRNA to the ribosome [27,28].

Of the 85 EST sequences representing *HBS1L*, 76 confirmed the existence of the previously identified primary transcript (NM_006620) [26]. Furthermore, our analysis of EST data revealed the use of three alternate polyadenylation signals, of which the middle signal would appear to be primarily utilized (producing a longer transcript than the reference transcript NM_006620). This is evidenced by both EST data and the 3 kb product size observed by Northern by both ourselves (data not shown) and others [28]. The full 7162 bp cDNA, with all three poly(A) recognition sites annotated has been submitted (AJ459826).

EST analyses followed by EST re-sequencing (as described in the later section) and RT-PCR revealed the existence of an exon 4A alternate splice variant, which contained precisely the same four first exons as the major, published *HBS1L* transcript. However, the transcript utilizes an alternate fifth exon "exon 4A", in which it terminates (missing the subsequent exons 5–18 of the primary transcript). This results in a 2800 bp cDNA (**AJ459827**), with the putative translation product encoding a 667AA protein. The exon 4A sequence contains an open reading frame resulting in an additional 489AA which are unique to this splice variant. However, the exon 4A protein sequence is novel with no significant homologies, and its function is entirely unknown. The structure of the *HBS1L* gene is detailed in Figure 3C.

ALDH8A1

ALDH8A1 was originally identified using a functional approach to identify genes responsible for 9-*cis*-retinal metabolism [29]. The *ALDH8A1* gene produces 2551 bp mRNA expressed in a variety of tissues including the erythroid and bone marrow. *ALDH8A1* [29] is represented by a 2551 bp cDNA (**AF303134**), comprising 7 exons spread over ~30 kb. Although we found no evidence for alternate splicing of this gene, the publication of the *ALDH8A1* gene [29] discusses the existence of an alternate transcript which skips exon 6. The structure of *ALDH8A1* is detailed in Figure 3D.

PDE7B

PDE7B [30,31] localizes to the distal (telomeric) extremity of the candidate interval, with only the promoter and first exon of the gene localized within the candidate interval. Due to this, extensive characterization of the gene was not performed and the gene has been thoroughly described elsewhere [30,31].

AHI1

The novel, previously uncharacterized gene *AHI1* was originally identified through numerous EST homologies to the corresponding genomic interval. EST data revealed the presence of 7 transcripts, two of which were experimentally verifiable by RT-PCR. The primary transcript (**AJ459824**) is a 5538 bp cDNA encoded by 28 exons spread over nearly 215 kb of DNA around the center of the candidate interval. This represents a huge gene (the average genomic extent is around 14 kb). A predicted CpG island exists which overlaps exon 1 of the gene. The encoded polypeptide is a 1197 amino acid protein with a predicted molecular weight of ~136 kDa. It contains six G-protein WD40 repeats in the region of exons 13–18 and a Src-homology 3 (SH3) domain in the region of exon 23 and 24. The N-terminus of the protein is novel, with the only significant homolog being an uncharacterized mouse gene (**BAB24355**). The existence of WD40 repeats and an SH3 domain suggests that **AJ459824** interacts with other proteins, possibly in the formation of large, transient multiprotein complexes. The diverse cellular roles of SH3 and WD-repeats precludes functional assignment of this gene.

The alternate splice variant 2 (**AJ459825**), which is 3654 bp, includes an alternate exon "22A" in which the transcript terminates (Figure 3B). This results in a protein containing the G-protein WD40 repeats, but missing the SH3 domain, suggesting that the two domains may function independently. Furthermore, RT-PCR and RACE investigation revealed a third gene variant (**AJ606362**). This was identical to the full-length transcript (**AJ459824**) in every respect except that it skipped exon 2. Further to the alternate splice variants, EST data revealed the use of an alternate polyadenylation signal on the major transcript (**AJ459824**).

The expression of *AHI1* was investigated experimentally and electronically. The gene would appear to be widely expressed, with detectable expression in all tissues tested by RT-PCR (small intestine, spleen, bone marrow, thymus, colon, testes and liver) except fetal liver (data not shown). "Digital differential display" generated from the source tissue of representative EST sequences [32], confirmed our experimental data, that this gene is widely expressed in kidney, germ cell, prostate, testis, uterus, whole embryo, brain, stomach, colon, pituitary, cervix, ovary, ear, eye, and lung.

EST based gene finding

EST re-sequencing was employed as a strategy for garnering further information concerning putative transcripts. As EST sequence generally represents a small part of the entire EST clone sequence (around 200–500 bp of publicly available sequence data, compared to clone lengths of up to and over 2 Kb), the full EST insert can be sequenced to extend the available sequence for each putative transcript. This strategy can therefore obtain the maximum quantity of information from an existing, cheap and readily available resource.

Strong EST homologies (greater than 95% identity over >100 bp) were identified by BLAST search [33] against the candidate interval sequence. By reference to the UniGene database [34], non-redundant clusters of ESTs were identified. These ESTs were assembled into an electronic contig using the contig assembly program Sequencher (GeneCodes). From each cluster of ESTs, the most appropriate clone (the most protruding, large insert clone) from each unmapped end of the cluster was selected and ordered from the IMAGE (Integrated Molecular Analysis of Genomes and their Expression) consortium [35]. All potentially useful EST sequences – which could provide novel sequence information – were then sequenced.

In total, approximately 13,000 EST sequences were found to be highly homologous to the candidate interval. However, the vast majority of these represented homology to the five pseudogenes within the candidate interval (briefly discussed later) and could be immediately discounted from providing useful information. A further five EST clusters represented the genes *MYB*, *HBS1L*, *ALDH8A1*, *PDE7B* and *AHI1*, and the results of their respective EST analysis were included earlier with reference to each gene.

After subtracting the known genes and the pseudogenes, a total of approximately 500 ESTs localized to the candidate interval, represented by 100 EST clusters. Of these 500 ESTs, 104 were deemed suitable for full-insert sequencing (based on insert size and position in EST contig; listed in table 3 (see additional file 1). Of the 104 sequenced ESTs, 7 contained a wrong, unexpected insert. The remaining 97 clones were sequenced to produce a total of 17 Kb of *novel* sequence data (i.e. potential transcript sequence data which was not previously available in any public database repository), confirming that EST re-sequencing is a viable method for generating novel transcript information. Furthermore, in six instances the novel EST sequence data revealed previously unidentified exons of a transcript.

Due to the large numbers of artifacts present in EST libraries (e.g. partially spliced RNA species, chimeras and contamination from various sources including genomic DNA [36-39]), we concentrated efforts on those EST transcripts that both exhibited exon structure and were completely homologous to the candidate interval. The uncharacterized EST clusters chiefly represented probable genomic contamination, with a smaller number representing chimeras.

This revealed 16 spliced, non-chimeric EST clusters, which were tested for expression by RT-PCR against a panel of cDNA tissues (fetal liver, erythroid, Lymphoblastoid cell line, spleen, stomach, thymus, small intestine, colon, bone marrow, testes and plasmid DNA from the cognate EST as a positive control). RT-PCR primers were designed such that the PCR product spanned an exon boundary, so that amplification from any contaminating genomic DNA could be eliminated. This experimentally verified the expression of four EST clusters.

To obtain full-length transcripts for these experimentally verified genes, we utilized the Clontech SMART-RACE system (primers sequences listed in the materials and methods). This produced full-length transcript information on the four genes, and revealed 11 splice variants of these genes. Table 3 lists the sizes, accession numbers and cDNA tissues that revealed detectable expression of these transcripts.

Analysis of these EST transcripts revealed that none contained a full-length open reading frame (see table 2), and there were no significant homologies to known genes. Therefore it is suggested that these transcripts either encode functional RNAs or represent aberrant transcription. We should add that the evidence of many of these

Table 2: Experimentally verified genes and transcripts identified through EST based evidence.

Gene	Transcript	Acc. No.	Length	Expression verified in:	Number of exons	Longest ORF
1	1	AJ606314	831 bp	Testes	6	138 bp
	2	AJ606325	717 bp	Testes	5	81 bp
	3	AJ606326	636 bp	Testes	4	150 bp
	4	AJ606327	522 bp	Testes	5	150 bp
2	1	AJ606328	2017 bp	Testes, Bone Marrow	5	294 bp
	2	AJ606329	1553 bp	Testes, Bone Marrow	3	294 bp
	3	AJ606330	1595 bp	Testes, Bone Marrow	5	294 bp
3	1	AJ606331	2501 bp	Testes, Placenta	10	258 bp
4	1	AJ606332	2106 bp	Testes	3	213 bp
	2	AJ606315	2390 bp	Testes	5	150 bp
	3	AJ606316	2718 bp	Testes, Placenta	3	210 bp

The genes have been named 1–4 for the purposes of this paper, with each gene being subdivided into separate transcripts (e.g. splice variants). The "Expression verified in" column lists which tissues we have verified the expression of the transcripts by RT-PCR (from a cDNA panel including Fetal Liver, Erythroid, Lymphoblastoid cell line, Spleen, Stomach, Thymus, Small Intestine, Colon, Bone Marrow, Placenta and Testes). The final column lists the longest Open Reading Frame identified in each cDNA, defined as the longest uninterrupted stretch between a methionine and a stop codon.

transcripts was purely derived from testes, a tissue previously associated with unusual transcripts that appear to be non-functional or aberrant versions of functioning genes [40,41]. The fact that a number of recent publications [42,43] have described testis-specific transcripts (some lacking ORFs) without mentioning the possibility of aberrant transcripts suggests that this area is worthy of further investigation and dissemination by the scientific community.

A further possibility is that non-coding RNA gene 4 is an antisense transcript of *PDE7B*, due to its location in the reverse orientation of intron 1 of *PDE7B*. However, it is more likely to have an unrelated function due to the non-overlapping intron/exon structure.

Comparative genomics

A novel comparative genomic strategy was used in an attempt to identify any genes which may have been missed using conventional approaches. PipMaker [44] was used to compare the candidate interval with the syntenic mouse region (on chromosome 10); all regions that had retained over 70% homology with the mouse within a sequence of 100 bp in length, but did not overlap the known genes, were identified. This homology cut-off has previously been deemed acceptable [45,46] and in-house analysis revealed that this level had a very high sensitivity for the known genes in the candidate interval, whilst not detecting a large number of surplus homologies. We based our gene finding strategy on the assumption that some of these human-mouse homologies may represent exons of undiscovered genes. Although a significant proportion of these homologies would be expected to represent regulatory elements, it is impossible to hypothesize which of the conserved blocks are expressed or regulatory. Therefore an inclusive strategy was designed that tested all homologies for expression without prior assumptions.

At the given cut-off, a total of 271 homologies were detected, with 59 overlapping known exons (~22%) and 68 located within introns of genes (~25%). A total of 144 homologies did not overlap known genes. Working on the assumption that some of these novel homologies may overlap undiscovered exons, we subjected the sequence of these homologous regions to the GeneSplicer [45] program. This predicted any potential splice sites within these putative exons, and thus delimited the range of the putative homologous exons. These predicted exons were then tested for expression using an RT-PCR based strategy. RNA from a panel of tissues (lymphoblastoid cell line, erythroid, bone marrow, spleen, fetal liver, liver and testes) was first DNAseI treated prior to reverse transcription. The cDNA was shown to be free from residual DNA contamination by the negative results of PCR with a panel of STS across the candidate interval (data not shown).

Primer pairs for each putative homologous exon (designed within the predicted splice sites; table 4 (see additional file 3)) were then tested for expression against the cDNA panel by RT-PCR, with genomic DNA as a positive control. Conserved regions that mapped within the extent of any of the known genes were not subject to RT-PCR. This was because such RT-PCR experiments could give a positive result due to the presence of partially processed heterogeneous nuclear RNA (i.e. unspliced intronic sequence). Therefore the outlined strategy does not have the power to identify putative genes that are localized within the introns of the known genes.

Figure 5 (see additional file 2) provides a PIP-plot of the candidate interval, annotated with the positions of genes and repetitive sequences, and overlaid with a summary of the experimental results. The detailed experimental results of this analysis are presented in Figure 4. The data revealed the positive expression of 12 conserved regions and weak positive PCR results for another 38 conserved regions, out of a total of 144 tested homologous regions. Several of these positive results (in the region of homology numbers 2→8) appeared to be from the same small region (<100 kb) at the centromeric side of the candidate interval and all expressed in the spleen. The other positive results were distributed across the candidate interval in no clear pattern, with the only other notable result being homology 233 which was positive in all tested tissues. To further investigate the existence of a gene expressed in the spleen at the left-hand side of the candidate interval, we attempted amplification spanning the relevant expressed homologous regions. This involved using e.g. the forward primer of homology 2 versus the reverse primer of homology 3, to see if a PCR product could be amplified, suggesting that these homologies were separate exons of the same gene. All such combinations of PCR primers positive from the comparative analysis were performed in an attempt to elucidate any possible gene structure. However, all PCRs were negative (data not shown), therefore the existence of such a gene is questionable. Additionally, the existence of the other positive results spread over the candidate interval remains unexplained. All these positive results could represent intergenic transcription [47,48]. Furthermore, the majority of the conserved sequences were confirmed as not being expressed in the cDNA panel, suggesting that their functional conservation may be related to control of gene expression. However, the cDNA panel was biased towards tissues of relevance to this study (i.e. red cell haematology) and it is therefore possible that some of these homologies may represent genes expressed in untested tissues.

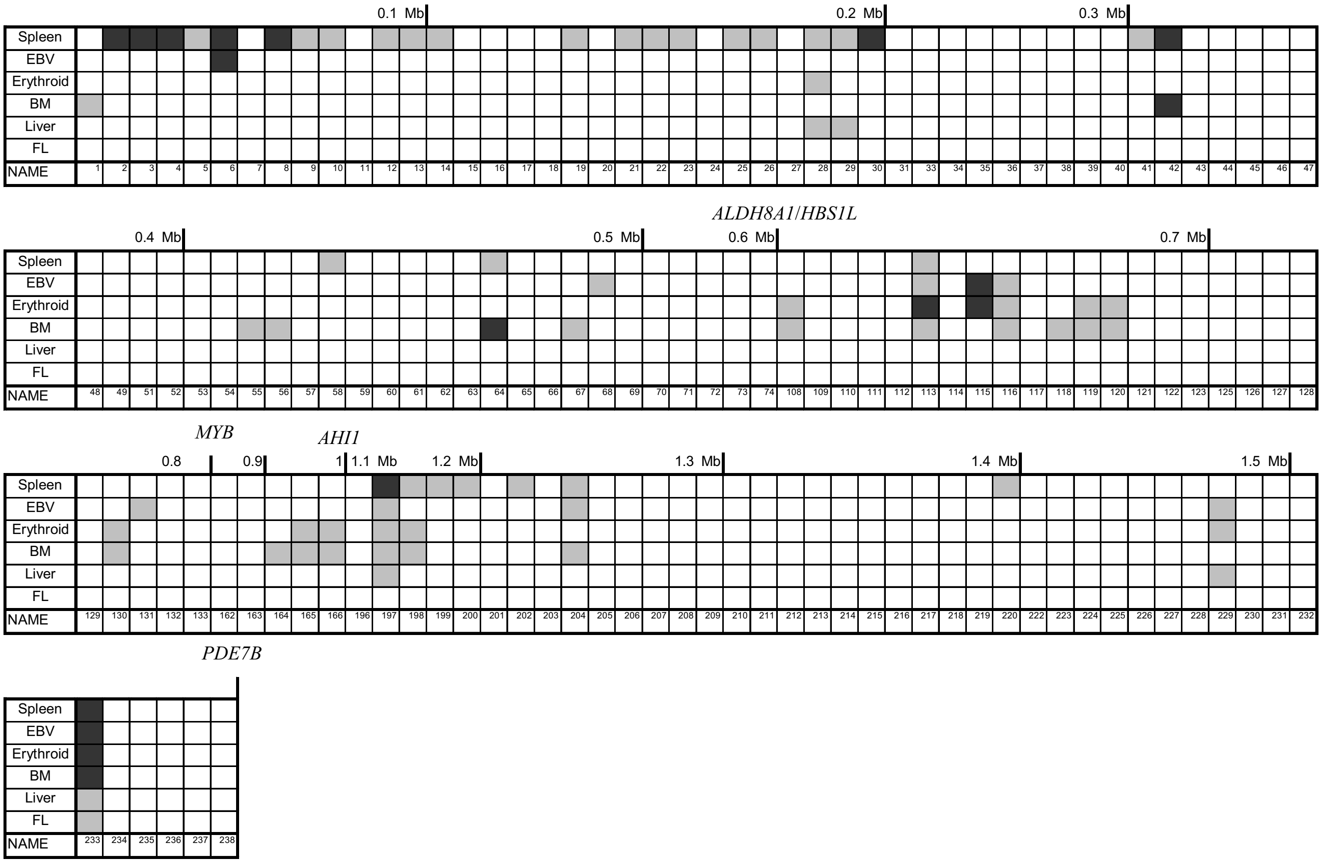

Figure 4

Results of comparative genomic analysis. Each column of the figure represents a region of homology from the 6q23 candidate interval with the mouse syntenic region (over 70% homology and 100 bp). These putative exon sequences are numbered from 1 to 238 at the bottom of each column. Conserved sequences corresponding to regions of a known gene were not included in the analysis, accounting for the missing numbers. As the conserved regions are not spread evenly across the candidate interval, the scale is variable but indicated above the table every 0.1 Mb. The locations of the known genes are also indicated above the table. The results of RT-PCR against the panel of cDNA tissues are indicated by black blocks for positive and gray block for weakly positive. Each row represents the results for each tissue and is indicated at the left: Spleen, EBV = Lymphoblastoid cell line, Erythroid, BM = Bone Marrow, Liver, FL = Fetal Liver.

Pseudogenes

All pseudogenes within the candidate interval represented processed pseudogenes, which are thought to arise by genomic integration of cDNA sequences generated by reverse transcription of an RNA transcript. Such pseudogenes are generally not transcribed because of a lack of functional promoters or other regulatory elements.

Five processed pseudogenes exist within the candidate interval – *GAPD* (NM_002046), *HMGA1* (NM_002128), *CGI-27* (NM_016139), *LOC51142* (NM_016139) and *FAM8A1* (NM_016255), indicated in Figure 1. The combined evidence of a disrupted ORF and lack of transcription provides substantial verification that the pseudogenes are not being transcribed. It is worth noting, however, that a central exon (exon 4) of RNA gene 1 has a small overlap (~100 bp) with HMGA1 pseudogene. The functional significance of this is not known.

Discussion

Nine genes were discovered in a region encompassing approximately 1.5 Mb of chromosome 6q23 using an integrated *in silico* and 'wet biology' approach with comparative genomics. These 9 genes produced a total of 30 different transcripts (around three different transcripts per gene) which is much higher than recent estimates of alternative splicing [49], a reflection perhaps, of the thorough annotation strategy that was used here.

The majority of the discovered alternate transcripts are unlikely to be functional at the protein level. This is due to the alternate exons interrupting and truncating the open reading frame. Of the 19 alternate splice variants of the 5 protein coding genes, only *MYB* exon 9A, *MYB* exon 8', *HBS1L* exon 4A, *AHI1* exon 22A and the *AHI1* transcript that skips exon 2 contain either further novel protein coding information or a complete open reading frame, suggesting that these splice variants are also protein coding. Truncation of the ORF in the majority of splice variants (particularly *MYB*) does not necessarily exclude a functional role for these alternative variants. For instance, the RNA could produce a shorter protein product, or could be involved in the control of gene expression. Recent evidence suggests that at least one third of alternative splice variants produce transcripts with premature stop codons, and that these transcripts may be involved with regulated unproductive splicing and translation (RUST). This mechanism of coupling of alternative splicing and nonsense mediated decay (NMD) may be a common and underappreciated means of regulating protein expression [50].

Further to the questionable nature of these alternate splice variants, there were four other non-coding RNA genes discovered in the candidate interval. The prevalence and importance of such transcripts has been grossly underestimated until recently and such genes would appear to be more common in genomic regions subjected to thorough annotation [51-53]. In particular, the most powerful approaches currently available (tiling paths of oligonucleotide microarrays) have revealed that around half of transcription occurs outside of any annotated genes, with such transcripts having lower and more limited expression. A large proportion of these transcripts appear to be non protein-coding, corroborating our findings [54-56]. Such results suggest that even thorough annotation may not uncover all transcriptional activity.

The short ORF of the non-coding transcripts does not preclude the hypothesis that these genes could in fact be protein coding, due to the existence of genes encoding very short proteins [57]. To question the validity of these genes further, we overlaid the exon sequence of these transcripts with the comparative mouse data. For all the non-coding transcripts, only a single exon of transcript 3 contained 70% identity over 100 bp with mouse sequence (homology 32 in Figure 5). Furthermore, these non-coding sequences were compared to the non-human EST databases by BLAST search to identify if any homologous genes existed in other species. Again, the only significant homologies were identified for transcript 3. This included two homologies with rat EST sequences (AA997847; 85% identity over 227 bp, and AI235855; 82% identity over 227 bp) and one homology with a mouse EST sequence (BY735461; 86% identity over 227 bp). This is strong evidence that sequence from transcript 3 is biologically relevant and conserved through evolution. However, the homologous ESTs represent anonymous, short sequences that are entirely uncharacterized and do not contain long open reading frames. Despite strong evidence of the functional nature of at least one of these apparently non-coding sequences, this lack of information precludes any functional assignment.

A range of techniques had been used to generate the transcripts in this region, each of which had various merits. We found resources such as ensembl, UCSC genome server and the NCBI map viewer invaluable, each providing a good overview of biologically relevant data within the candidate interval. However, whilst such data is rarely totally incorrect, it is error prone and requires manual inspection and confirmation. The homology and EST data provided a second tier of valuable information. Whilst much of this information overlapped known genes, both sources recognized apparently functional alternate transcripts and novel genes that would have otherwise been missed. Because the annotations based on these approaches are biased strongly in favor of protein-coding transcripts, a large proportion of non-coding RNAs could still be missed, as demonstrated by recent studies using high-density oligonucleotide arrays [54-56].

The detailed transcript map is presently being used as a guide for mutation analyses in the Asian-Indian kindred [13], with the aim of identifying the specific genetic lesion responsible for raised fetal hemoglobin expression. Whilst we are making no assumptions concerning the gene responsible, it is possible to use functional and bioinformatic information for prioritisation. In particular, *MYB*, a haematopoietic-specific transcription factor, is known to influence proliferation, differentiation and cell cycle progression and therefore presents itself as the most obvious candidate. However, the presence of SH3 and WD40 repeats in *AHI1* suggests that it is involved with protein-protein interactions and could, potentially influence a diversity of cellular functions, therefore it remains a good candidate gene. The other three protein-coding genes appear to be the least viable candidates – *PDE7B* contains only one exon within the candidate interval, *ALDH8A1* is not expressed in the candidate tissue (erythroid cells) and *HBS1L* is a translation elongation factor. We are currently in the process of analysing all the protein coding genes for mutations and have initiated *in-vitro* cellular studies to examine protein function.

Methods

Bioinformatic analysis of candidate interval

All described analysis were performed on the NCBI assembly 29 of human chromosome 6, with the candidate region located to 146786098-148413740 bp of this assembly. The syntenic mouse region (for PIP analysis) comprised the reverse complement of chromosome 10: 20172930-21905929 bp on the MGSCv3 build. Initial analysis of the genomic region included analysis of GC content using the EMBOSS [58] program "geecee" and analysis of repeats using RepeatMasker [59]. BLAST searches were performed using default parameters for nucleotide sequences against the human nucleotide database [33] and the miRNA registry [60]. For the identification of human EST homologies, BLAST searchers were performed using default parameters against the human EST database. Strong EST homologies (greater than 95% identity over >100 bp) were identified by parsing the results using an in-house Perl script.

Annotation of individual candidate genes was achieved using a combination of Ensembl [61] and the UCSC genome browser [62], with the gene predictions in table 1 being derived from these genome browsers.

PCR amplification of short products

PCR reactions were of a total volume of 15 μl or 50 μl, with a template of approximately 50 ng genomic DNA, 1 ng Plasmid DNA or 50 ng cDNA, 5–10 pMoles of the forward and reverse PCR primers, 0.1 mM of each dNTP (Boehringer Mannheim), 1.5–2.5 mM $MgCl_2$, and 1U of TaqGold DNA polymerase (PE Biosciences). Typical thermal cycling parameters consisted of an initial denaturation step of 94°C for 10 mins, followed by 30–35 cycles of denaturation (94°C for 1 min), annealing (50°–60°C for 1 min) and extension (72°C for 1 min), and a final extension of 72°C for 5 mins. 1–10 μl of PCR product was resolved on a 0.8–2% agarose gel containing Ethidium Bromide (EtBr) by electrophoresis.

RNA samples

The following human total RNA samples were purchased from Clontech: Bone marrow, colon, small intestine, spleen, stomach, thymus, testes and human fetal liver. Placental RNA was included with the clontech SMART-RACE kit (described later). Erythroid progenitor RNA was obtained using the Fibach culture [63] and B-cell RNA from EBV cell lines using standard methodologies. RNA was prepared by guanidinium-thiocynate phenol-chloroform extraction [64].

Synthesis of first-strand cDNA

Reverse transcription (RT) reactions were performed by initially mixing 1 μg of RNA, 0.5 μg oligo(dT) primer *or* 0.2 μg random hexamer in a total of 10 μl, incubating at 70°C for 5' and chilling on ice. The following components were then added; 4 μl of 5 × reaction buffer, 2 μl 10 mM dNTP mix, 1 μl (20 u) ribonuclease inhibitor and nuclease free deionized water to 19 μl. The random hexamer reaction was incubated at 25°C for 5 min and the Oligo(dT) reaction incubated at 37°C for 5 min. 200U (1 μl) of SuperScript II (GibcoBRL) was added to each reaction, followed by incubation at the appropriate temperature for the enzyme for 1 hour. The reaction was stopped at 70°C for 10 mins followed by pooling of the random hexamer and oligo(dT) reactions of the same sample. For RT-PCR, 1 μl of first strand cDNA was utilized as template in standard PCR reactions.

Sequencing IMAGE EST clones

Plasmid DNA for IMAGE clones was prepared using the Qiagen plasmid mini-prep spin kit according to the manufacturer's conditions, and sequencing was performed using Perkin Elmer BigDye terminator chemistry under standard conditions.

RACE

All RACE reactions were carried out using the Clontech SMART RACE cDNA Amplification Kit, according to the Clontech manufacturer's guidelines. These RACE products were gel purified and cloned directly using the TOPO TA type PCR kit for sequencing (Invitrogen life technologies), prior to sequencing of cloned inserts by BigDye terminator chemistry (Applied Biosystems). For some transcript sequences the 3' end of the transcript was already known due to the presence of poly(A) tails and signals, so only 5' RACE was required. In the case of transcript 1, a second, nested RACE primer was used to get the full-length transcript. The primers used for RACE extension of transcript sequences are as follows:

Non-coding transcript 1-5'-RACE: TGTGTAGTCAAGTCCAGTCATCAGCAG, 1-5'RACE2: AGACTGTTCCCGCTCGCGTGGCCG, Non-coding transcript 2-3'-RACE: TATTGACACGGCGCTACCACGGG, Non-coding transcript 2-5'-RACE: CATCATCTTCCCCTGTTTGTTAGCTT, Non-coding transcript 3-3'-RACE: ATGGAACAGACTAGAAGGAC, Non-coding transcript 3-5'-RACE: AGTGCCACCTGACCGCTTGA, Non-coding transcript 4-5'-RACE: GCAAGGTGGTGTCAACATGGTAGG

Comparative genomic analysis

Initial comparative genomic analysis was performed using PipMaker [44], with the following syntenic sequences used for analysis: 1) human chromosome 6:146786098-148413740 bp from NCBI assembly 29 and 2) mouse chromosome 10:20172930-21905929 bp from MGSCv3 and the PipMaker options "search one strand" and "chaining". The "concise" output file from PipMaker was parsed with in-house Perl scripts to select all homologies with >70% homology over 100 bp, to remove all homologies which overlapped with known genes in the candidate interval, and to extract DNA sequences for each of these selected homologous sequences. These sequences were analyzed with GeneSplicer [65] (with options set at -e 1 -a 1), and the region on which RT-PCR primers were designed reduced to within the best predicted splice site. All primers were then designed by eprimer3 for RT-PCR (table 4 (see additional file 3)).

RNA for RT-PCR was first subjected to DNAseI digestion (GibcoBRL) according to manufacturer's conditions. 1st strand synthesis and RT-PCR was performed as described earlier.

Accession Numbers

Sequence data from this article have been deposited with the DDBJ/EMBL/GenBank Libraries [66] under the following Accession Numbers: AJ459824 (*AHI1* primary transcript); AJ459825 (*AHI1* exon 22A alternate splice variant); AJ606362 (*AHI1* splice variant skips exon 2); AJ459826 (*HBS1L* primary transcript); AJ459827 (*HBS1L* exon 4A alternate splice variant); AJ606317 (*MYB* exon 8A splice variant); AJ606318 (*MYB* exon 9Ai splice variant); AJ606319 (*MYB* exon 9Aii splice variant); AJ606320 (*MYB* exon 8' splice variant); AJ606321 (*MYB* exon 8' and 8A splice variant); AJ606322 (*MYB* exon 10A splice variant); AJ606323 (*MYB* exon 13A splice variant); AJ606324 (*MYB* exon 14A splice variant); AJ606314 (Non-coding transcript 1-1); AJ606325 (Non-coding transcript 1-2); AJ606326 (Non-coding transcript 1-3); AJ606327 (Non-coding transcript 1-4); AJ606328 (Non-coding transcript 2-1); AJ606329 (Non-coding transcript 2-2); AJ606330 (Non-coding transcript 2-3); AJ606331 (Non-coding transcript 3-1); AJ606332 (Non-coding transcript 4-1); AJ606315 (Non-coding transcript 4-2); AJ606316 (Non-coding transcript 4-3); AJ606363 (microsatellite marker DbAD6)

Authors' contributions

JC carried out molecular genetic, bioinformatic and sequence alignment studies, and drafted the manuscript; LG participated in the molecular genetic studies; BC participated in the genomic and sequence alignment studies; JB participated in the molecular genetic studies; AG participated in PCR and sequencing studies. SLT conceived of the study, and participated in its design and co-ordination. All authors read and approved the final manuscript.

Additional material

Additional File 1

Table 3. Table showing NCBI accession numbers of ESTs ordered and sequenced from the IMAGE consortium.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-33-S1.doc]

Additional File 3

Table 4. Primers used for testing expression of conserved regions from comparative genomic analysis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-33-S3.doc]

Additional File 2

Figure 5. PIP-plot of the candidate interval, annotated with the positions of genes and repetitive sequences, and overlaid with a summary of the experimental results. Pips which were found to be experimentally expressed are highlighted by green bars, weak positives are indicated by blue bars and negative pips are highlighted by red bars. Results of comparative genomics in PIP-plot format. This figure provides the raw graphic output of the PipMaker program. This is annotated with the primary transcripts of the protein coding genes and all the exons of the non-coding RNA transcripts. Alternate transcripts and pseudogenes are not detailed due to limitations with the PipMaker program display. Therefore only the primary transcript of the protein coding genes is displayed, and all exons of the non-coding RNA genes are displayed (as the major transcript is unknown). The PIP-plot is overlaid with the experimental results testing the expression of the individual PIPs. This is detailed as colored vertical bars with red bars representing negative (i.e. not expressed) PIPs, blue representing weakly positive PIPs and green bars representing positive (i.e. expressed) PIPs. The positions of repeat sequences are indicated on the figure.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-33-S2.pdf]

Acknowledgments

We thank Claire Steward for preparation of the manuscript; Dalla-Favera, R., Wallrap, C. and Lin., M, for kindly providing cDNA clones of candidate genes (*MYB*, *HBS1L* and *ALDH8A1*, respectively). This work was supported by the Medical Research Council, UK (Grant G0000111; ID51640) to SLT. JB was supported by a study visit grant from the Royal Society, UK.

References

1. Weatherall DJ, Clegg JB: **Thalassemia - a global public health problem.** *Nature Medicine* 1996, **2:**847-849.
2. Beutler E: **Discrepancies between genotype and phenotype in hematology: an important frontier.** *Blood* 2001, **98:**2597-2602.
3. Platt OS, Brambilla DJ, Rosse WF, Milner PF, Castro O, Steinberg MH, Klug PP: **Mortality in sickle cell disease: Life expectancy and risk factors for early death.** *New England Journal of Medicine* 1994, **330:**1639-1644.
4. Ho PJ, Hall GW, Luo LY, Weatherall DJ, Thein SL: **Beta thalassemia intermedia: is it possible to consistently predict phenotype from genotype?** *British Journal of Haematology* 1998, **100:**70-78.
5. Boyer SH, Belding TK, Margolet L, Noyes AN: **Fetal hemoglobin restriction to a few erythrocytes (F cells) in normal human adults.** *Science* 1975, **188:**361-363.

6. Zago MA, Wood WG, Clegg JB, Weatherall DJ, O'Sullivan M, Gunson H: **Genetic control of F cells in human adults.** *Blood* 1979, **53:**977-986.
7. Garner C, Tatu T, Reittie JE, Littlewood T, Darley J, Cervino S, Farrall M, Kelly P, Spector TD, Thein SL: **Genetic influences on F cells and other hematologic variables: a twin heritability study.** *Blood* 2000, **95:**342-346.
8. Rutland PC, Pembrey ME, Davies T: **The estimation of fetal haemoglobin in healthy adults by radioimmunoassay.** *British Journal of Haematology* 1983, **53:**673-682.
9. Miyoshi K, Kaneto Y, Kawai H, Ohchi H, Niki S, Hasegawa K, Shirakami A, Yamano T: **X-linked dominant control of F-cells in normal adult life: Characterization of the Swiss type as hereditary persistence of fetal hemoglobin regulated dominantly by gene(s) on X chromosome.** *Blood* 1988, **72:**1854-1860.
10. Gilman JG, Huisman THJ: **DNA sequence variation associated with elevated fetal Gg globin production.** *Blood* 1985, **66:**783-787.
11. Garner C, Tatu T, Game L, Cardon LR, Spector TD, Farrall M, Thein SL: **A candidate gene study of F cell levels in sibling pairs using a joint linkage and association analysis.** *GeneScreen* 2000, **1:**9-14.
12. Wood WG: **Hereditary Persistence of Fetal Hemoglobin and Delta Beta Thalassemia.** *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management* Edited by: SteinbergMH, ForgetBG, HiggsDR and NagelRL. Cambridge, UK, Cambridge University Press; 2001:356-388.
13. Thein SL, Sampietro M, Rohde K, Rochette J, Weatherall DJ, Lathrop GM, Demenais F: **Detection of a major gene for heterocellular hereditary persistence of fetal hemoglobin after accounting for genetic modifiers.** *American Journal of Human Genetics* 1994, **54:**214-228.
14. Craig JE, Rochette J, Fisher CA, Weatherall DJ, Marc S, Lathrop GM, Demenais F, Thein SL: **Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach.** *Nature Genetics* 1996, **12:**58-64.
15. Garner C, Mitchell J, Hatzis T, Reittie J, Farrell M, Thein SL: **Haplotype mapping of a major QTL for fetal hemoglobin production on chromosome 6q23.** *American Journal of Human Genetics* 1998, **62:**1468-1474.
16. Game L, Close J, Stephens P, Mitchell J, Best S, Rochette J, Louis-dit-Sully C, Riley J, See CG, Sanseau P, Kearney L, Bethel G, Humphray S, Dunham I, Mungall A, Thein SL: **An integrated map of human 6q22.3-q24 including a 3 Mb high resolution BAC/PAC contig encompassing a QTL for fetal hemoglobin.** *Genomics* 2000, **64:**264-276.
17. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan: **Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium.** *Nature* 2001, **409:**860-921.
18. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25:**955-964.
19. Oh IH, Reddy EP: **The myb gene family in cell growth, differentiation and apoptosis.** *Oncogene* 1999, **18:**3017-3033.
20. Majello B, Kenyon LC, Dalla-Favera R: **Human c-myb protooncogene: Nucleotide sequence of cDNA and organization of the genomic locus.** *Proc Natl Acad Sci U S A* 1986, **83:**9636-9640.
21. Dasgupta P, Reddy EP: **Identification of alternatively spliced transcripts for human c-myb: Molecular cloning and sequence analysis of human c-myb: exon 9A sequences.** *Oncogene* 1989, **4:**1419-1423.
22. Westin EH, Gorse KM, Clarke MF: **Alternative splicing of the human c-myb gene.** *Oncogene* 1990, **5:**1117-1124.
23. Slamon DJ, Boone TC, Murdock DC, Keith DE, Press MF, Larson RA, Souza LM: **Studies of the human c-myb gene and its product in human acute leukemias.** *Science* 1986, **233:**347-351.
24. Dudek H, Reddy EP: **Identification of two translational products for c-myb.** *Oncogene* 1989, **4:**1061-1066.
25. Shen-Ong GL: **Alternative internal splicing in c-myb RNAs occurs commonly in normal and tumor cells.** *Embo J* 1987, **6:**4035-4039.
26. Wallrapp C,, Müller-Pillasch F, Solinas-Toldo S, Lichter P, Friess H, Büchler M, Fink T, Adler G, Gress TM: **Characterization of a high copy number amplification at 6q24 in pancreatic cancer identifies c-myb as a candidate oncogene.** *Cancer Research* 1997, **57:**3135-3139.
27. Nelson RJ, Ziegelhoffer T, Nicolet C, Werner-Washburne M, Craig EA: **The translation machinery and 70 kd heat shock protein cooperate in protein synthesis.** *Cell* 1992, **71:**97-105.
28. Wallrapp C, Verrier S-B, Zhouravleva G, Philippe H, Philippe M, Gress TM, Jean-Jean O: **The product of the mammalian orthologue of the Saccharomyces cerevisiae HBS1 gene is phylogenetically related to eukaryotic release factor 3 (eRF3) but does not carry eRF3-like activity.** *FEBS Letters* 1998, **440:**387-392.
29. Lin M, Napoli JL: **cDNA cloning and expression of a human aldehyde dehydrogenase (ALDH) active with 9-cis-retinal and identification of a rat ortholog, ALDH12.** *J Biol Chem* 2000, **275:**40106-40112.
30. Sasaki T, Kotera J, Yuasa K, Omori K: **Identification of human PDE7B, a cAMP-specific phosphodiesterase.** *Biochem Biophys Res Commun* 2000, **271:**575-583.
31. Gardner C, Robas N, Cawkill D, Fidock M: **Cloning and characterization of the human and mouse PDE7B, a novel cAMP-specific cyclic nucleotide phosphodiesterase.** *Biochem Biophys Res Commun* 2000, **272:**186-192.
32. **SOURCE Search** [http://genome-www5.stanford.edu/cgi-bin/source/sourceSearch]
33. **NCBI BLAST** [http://www.ncbi.nlm.nih.gov/BLAST/]
34. **UniGene** [http://www.ncbi.nlm.nih.gov/UniGene]
35. **MRC Geneservice** [http://www.hgmp.mrc.ac.uk/Biology/descriptions/image.html]
36. Zhuo D, Zhao WD, Wright FA, Yang HY, Wang JP, Sears R, Baer T, Kwon DH, Gordon D, Gibbs S, Dai D, Yang Q, Spitzner J, Krahe R, Stredney D, Stutz A, Yuan B: **Assembly, annotation, and integration of UNIGENE clusters into the human genome draft.** *Genome Res* 2001, **11:**904-918.
37. Wolfsberg TG, Landsman D: **A comparison of expressed sequence tags (ESTs) to human genomic sequences.** *Nucleic Acids Research* 1997, **25:**1626-1632.
38. Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, Hawkins M, Hultman M, Kucaba T, Lacy M, Le M, Le N, Mardis E, Moore B, Morris M, Parsons J, Prange C, Rifkin L, Rohlfing T, Schellenberg K, Marra M, et al.: **Generation and analysis of 280,000 human expressed sequence tags.** *Genome Res* 1996, **6:**807-828.
39. Bailey L. C., Jr., Searls DB, Overton GC: **Analysis of EST-driven gene annotation in human genomic sequence.** *Genome Res* 1998, **8:**362-376.
40. Ivell R: **'All that glisters is not gold' - common testis gene transcripts are not always what they seem.** *Int J Androl* 1992, **15:**85-92.
41. Jeannotte L, Burbach JP, Drouin J: **Unusual proopiomelanocortin ribonucleic acids in extrapituitary tissues: intronless tran-**

scripts in testes and long poly(A) tails in hypothalamus. *Mol Endocrinol* 1987, **1:**749-757.

42. Braun A, Aszodi A, Hellebrand H, Berna A, Fassler R, Brandau O: **Genomic organization of profilin-III and evidence for a transcript expressed exclusively in testis.** *Gene* 2002, **283:**219-225.
43. Zendman AJ, van Kraats AA, den Hollander AI, Weidle UH, Ruiter DJ, van Muijen GN: **Characterization of XAGE-1b, a short major transcript of cancer/testis-associated gene XAGE-1, induced in melanoma metastasis.** *Int J Cancer* 2002, **97:**195-204.
44. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker--a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10:**577-586.
45. DeSilva U, Elnitski L, Idol JR, Doyle JL, Gan W, Thomas JW, Schwartz S, Dietrich NL, Beckstrom-Sternberg SM, McDowell JC, Blakesley RW, Bouffard GG, Thomas PJ, Touchman JW, Miller W, Green ED: **Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome.** *Genome Res* 2002, **12:**3-15.
46. Footz TK, Brinkman-Mills P, Banting GS, Maier SA, Riazi MA, Bridgland L, Hu S, Birren B, Minoshima S, Shimizu N, Pan H, Nguyen T, Fang F, Fu Y, Ray L, Wu H, Shaull S, Phan S, Yao Z, Chen F, Huan A, Hu P, Wang Q, Loh P, Qi S, Roe BA, McDermid HE: **Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere.** *Genome Res* 2001, **11:**1053-1070.
47. Plant KE, Routledge SJ, Proudfoot NJ: **Intergenic transcription in the human beta-globin gene cluster.** *Mol Cell Biol* 2001, **21:**6507-6514.
48. Ogawa Y, Lee JT: **Xite, X-inactivation intergenic transcription elements that regulate the probability of choice.** *Mol Cell* 2003, **11:**731-743.
49. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P: **Alternative splicing and genome complexity.** *Nat Genet* 2001, **30:**29-30.
50. Lewis BP, Green RE, Brenner SE: **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.** *Proc Natl Acad Sci U S A* 2003, **100:**189-192.
51. Eddy SR: **Non-coding RNA genes and the modern RNA world.** *Nat Rev Genet* 2001, **2:**919-929.
52. Meguro M, Mitsuya K, Nomura N, Kohda M, Kashiwagi A, Nishigaki R, Yoshioka H, Nakao M, Oishi M, Oshimura M: **Large-scale evaluation of imprinting status in the Prader-Willi syndrome region: an imprinted direct repeat cluster resembling small nucleolar RNA genes.** *Hum Mol Genet* 2001, **10:**383-394.
53. Cavaille J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie JP, Brosius J, Huttenhofer A: **Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization.** *Proc Natl Acad Sci U S A* 2000, **97:**14311-14316.
54. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SPA, Gingeras TR: **Large-Scale Transcriptional Activity in Chromosomes 21 and 22.** *Science* 2002, **296:**916-919.
55. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14:**331-342.
56. Cawley S, Bekiranov S, Ng HH, P Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, R. Gingeras T.: **Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs.** *Cell* 2004, **116:**499-509.
57. Gonzalez-Pastor JE, San Millan JL, Moreno F: **The smallest known gene.** *Nature* 1994, **369:**281.
58. **EMBOSS** [http://www.hgmp.mrc.ac.uk/Software/EMBOSS/]
59. **RepeatMasker** [http://www.repeatmasker.org/]
60. **The miRNA Registry** [http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml]
61. **Ensembl Genome Browser** [http://www.ensembl.org]
62. **UCSC Genome Browser** [http://genome.ucsc.edu]
63. Fibach E, Manor D, Oppenheim A, Rachmilewitz EA: **Proliferation and maturation of human erythroid progenitors in liquid culture.** *Blood* 1989, **73:**100-103.
64. Chomczynski P, Sacchi N: **Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.** *Analytical Biochemistry* 1987, **162:**156-159.
65. Pertea M, Lin X, Salzberg SL: **GeneSplicer: a new computational method for splice site prediction.** *Nucleic Acids Res* 2001, **29:**1185-1190.
66. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2003, **31:**23-27.