Global gene expression analysis of human erythroid progenitors

Alison T. Merryweather-Clarke\textsuperscript{1,2}, Ann Atzberger\textsuperscript{1*}, Shamit Soneji\textsuperscript{1}, Nicki Gray\textsuperscript{1}, Kevin Clark\textsuperscript{1}, Craig Waugh\textsuperscript{3}, Simon J. McGowan\textsuperscript{4}, Stephen Taylor\textsuperscript{4}, Asoke K. Nandi\textsuperscript{5}, William G. Wood\textsuperscript{1}, David J. Roberts\textsuperscript{2,6}, Douglas R. Higgs\textsuperscript{1}, Veronica J. Buckle\textsuperscript{1}, Kathryn J.H. Robson\textsuperscript{1}

\textsuperscript{1}MRC Molecular Hematology Unit, Weatherall Institute of Molecular Medicine, Headley Way, Headington, Oxford, OX3 9DS, UK; \textsuperscript{2}Blood Research Laboratory, Nuffield Department of Clinical Laboratory Sciences, Level 4, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK; \textsuperscript{3}MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, Headley Way, Headington, Oxford, OX3 9DS, UK; \textsuperscript{4}Computational Biology Research Group, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DS, UK; \textsuperscript{5}Department of Electrical Engineering and Electronics, The University of Liverpool, Brownlow Hill, Liverpool, L69 3GJ, UK; \textsuperscript{6}NHSBT – Oxford, Level 2, John Radcliffe Hospital, Headington, Oxford, OX3 9BQ, UK.

*Current address: Flow Cytometry Laboratory, Institute of Molecular Medicine Trinity College Dublin, Dublin 2 (Ireland)

Corresponding Author:
Dr Alison Merryweather-Clarke
Blood Research Laboratory,
Nuffield Department of Clinical Laboratory Sciences,
Level 4, John Radcliffe Hospital,
Headington,
Oxford,
OX3 9DU
e-mail : alison.merryweather-clarke@ndcls.ox.ac.uk
Tel : 44 1865 387907/387945
Fax : 44 1865 387957

Running title: Erythroblast transcriptome

Key words: Erythropoiesis, gene expression, microarray, red cell maturation

The "Human Erythroblast Maturation (HEM) Database" is available at:
https://cellline.molbiol.ox.ac.uk/eryth/index.html

The database is protected by a username/password (prior to publication). External users can log in using a guest account:

- username: auser
- password: erythEXPR
Abstract

Understanding the pattern of gene expression during erythropoiesis is crucial for a synthesis of erythroid developmental biology. Here, we isolated four distinct populations at successive erythropoietin-dependent stages of erythropoiesis including the terminal, pyknotic stage. The transcriptome was determined using Affymetrix arrays. First, we demonstrated the importance of using defined cell populations to identify lineage and temporally-specific patterns of gene expression. Cells sorted by surface expression profile not only express significantly fewer genes than unsorted cells, but also demonstrate significantly greater differences in the expression levels of particular genes between stages than unsorted cells. Second, using standard software we identified over 1000 transcripts not previously observed to be differentially expressed during erythroid maturation, 13 of which are highly significantly terminally regulated, including RFXAP and SMARCA4. Third, using matched filtering we identified 12 transcripts not previously reported to be continuously up-regulated in maturing human primary erythroblasts. Finally, using transcription factor binding site analysis we identified potential transcription factors that may regulate gene expression during terminal erythropoiesis. Our stringent lists of differentially regulated and continuously expressed transcripts containing many genes with undiscovered functions in erythroblasts are a resource for future functional studies of erythropoiesis. Our Human Erythroid Maturation database is available at https://cellline.molbiol.ox.ac.uk/eryth/index.html.
Introduction

The study of hematopoiesis has increased our understanding of stem cells and how they differentiate along the lymphoid and myeloid lineages in response to transcription factors (reviewed in 1-3). To date much of the interest has focused on the very early stages in this process, providing insight into those involved in differentiating from a stem cell to a committed progenitor (reviewed in 2,3). Erythropoiesis is a specialized process producing a cell dedicated to the delivery of oxygen to the tissues. As erythroblasts mature, they decrease in size, synthesise more hemoglobin and undergo membrane re-organization and chromatin condensation. Eventually the cells expel their organelles prior to taking up the biconcave discoid structure.

Understanding the pattern of gene expression and identifying the specific genes expressed during erythropoiesis is crucial for a synthesis of erythroid developmental biology and to define the molecular pathology of congenital and acquired dyserythropoietic syndromes.

Previous erythroblast microarray expression studies have focused on mouse cells or knockout models4-7 or used human cell lines8 or mixed populations of human erythroid cells of varying purity and maturity.9-13 We have cultured erythroblasts from peripheral blood mononucleocytes (PBMCs) and used flow cytometric cell sorting to isolate discrete populations of colony forming unit-Erythroid (CFU-E), pro-erythroblasts (Pro-E), intermediate (Int-E) and late (Late-E) erythroblasts.

Clustering transcripts using available proprietary software has a number of limitations; in particular, the lack of knowledge of the number of clusters, the large variability of changes in expression levels of biologically relevant co-regulated genes and the availability of measurements at a very few time points for each gene (in our dataset, this number is four). To address these issues, we have begun to explore other methods and here we present analysis of this data using matched filtering methods.

Transcription factors are integral to commitment;3 GATA1 is essential for erythroid maturation. GATA1-/- erythroblasts in vitro fail to mature beyond the Pro-E stage, suggesting that GATA1 is essential for erythroblast survival, while friend of GATA1 (FOG1) and Tal1 are also integral to erythropoiesis. A comprehensive description of the genes expressed during erythroid development has now allowed us to analyse transcription factor binding sites of erythroid-specific transcripts.

To elucidate the mechanisms involved in terminal red blood cell development, we have studied global gene expression patterns in human erythroid progenitors at several stages of differentiation using Affymetrix HGU133_plus_2.0 arrays. Here we present a comprehensive reference expression database of pure populations of normal human erythroblasts including several previously undescribed differentially regulated erythroid genes with unknown roles in erythropoiesis. Our data will facilitate studies of erythropoiesis in health and disease.
Materials and Methods

Cell culture

Human primary differentiating erythroblasts derived from peripheral blood buffy coat (NHSBT, Bristol, UK) were obtained with informed consent in accordance with the Declaration of Helsinki. All experiments were conducted with approval from the UK Home Office. Erythroblasts were cultured using a two-phase liquid culture system as previously described. In brief, PBMCs were obtained from peripheral blood buffy coat by Ficoll-Hypaque density-gradient centrifugation and seeded at 2.5 x 10^6 cells/ml in MEM with 10% FCS, 1 µg/ml cyclosporin A, and 10% conditioned medium from the 5637 bladder carcinoma cell line. The cells were cultured for 6–7 d (phase I), and the non-adherent cells were washed and re-seeded in fresh medium with 30% FCS, 1% deionized BSA, 1 U/ml of human recombinant erythropoietin (EPO), 10^5 M β-mercaptoethanol, 10^-6 mol/l dexamethasone, 0.3 mg/ml holotransferrin (ICN Biochemicals), and 10 ng/ml of human stem cell factor (phase II). Maturation of erythroblasts was monitored by May Grünwald-Giemsa stained cytospin preparations using light microscopy. Cells were harvested following phase II culture in the presence of EPO for four, five-six, eight-eleven or thirteen-fifteen days to obtain populations of CFU-Es, Pro-Es, Int-Es, or Late-Es, respectively.

Cell Sorting

Cells were stained and sorted using anti–CD36-PE (BD Biosciences; used for CFU-Es, Pro-Es and Int-Es), anti–CD71-FITC (DakoCytomation), anti–CD235a-APC (BD Biosciences; used for CFU-Es, Pro-Es and Int-Es), and anti–CD235a-RPE (DakoCytomation; used for Late-Es).

RNA Extraction

The purity and morphology of all sorted erythroblast populations was confirmed by cytospin. Total RNA was extracted using Trizol (Invitrogen, Paisley, UK) according to the manufacturer’s instructions, appending a phenol-chloroform step. RNA integrity was evaluated using an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA).

Biotinylated cRNA Preparation and Hybridization

For each of the four stages of erythroid precursors included in the study, three RNA pools were prepared from non-overlapping sets of four different RNA preparations mixed in equal proportions. One RNA pool of four RNA preparations from unsorted cells harvested at each of the four different stages of culture was also prepared for comparison. Biotinylated fragmented cRNA was synthesized from 100ng of pooled RNA using the 2-Cycle cDNA Synthesis and the 2-Cycle Target Labelling and Control Reagent kits (Affymetrix, Santa Clara, CA), and hybridized to
GeneChip Human Genome U133_Plus_2.0 arrays (Affymetrix) following the manufacturer’s recommendations.

**Analysis of Microarray Data**

Cell intensity calculation and scaling was performed using GeneChip Operating Software (GCOS) (Affymetrix) and data analyzed by the Robust Multiarray Average (RMA) method of Irizarry using the Bioconductor suite of packages in R (www.bioconductor.org). Data pre-processing and statistical analysis of differential expression in R used the Linear Models for Microarray Data (LIMMA) package from Bioconductor. The returned B-values (log odds of differential expression) and p-values were used to cluster the data.

Heat maps of RNA expression were generated using Genesis. DiRE, Pscan and CONFAC were used for transcription factor binding site (TFBS) discovery. Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com) and DAVID were used for ontology mapping.

Sets of normally distributed data were compared using Student’s t-tests in GraphPad Prism 5 (GraphPad Software Inc, San Diego, CA, USA).

To compare our data from Affymetrix HGU133_plus_2.0 with previously published data from Illumina HumanWG-6 v2 Expression BeadChips, we cross-referenced the unique probe identifiers (PIDs) using the IlluminaV2_AffyMapping.txt and HumanWG-6 v2 manifest files available at www.switchtoi.com and external gene names. 33966 Illumina PIDs matched annotated Affymetrix PIDs representing 15858 genes from the Illumina array. Of 7599 Illumina IDs present in erythroblasts, 7150 cross-matched to 6976 Affymetrix PIDs representing 6911 genes, represented on the Affymetrix HGU133_plus_2.0 chip by 16120 Affymetrix PIDs.


We have also retrospectively analyzed other datasets deposited with NCBIGEO containing expression measurements in erythroblasts at more than one time-point of erythropoiesis as indicated in Table 1, using RMA (where the CEL files are available) and LIMMA, for comparison with our data.

**Web-accessible database**

Following RMA and LIMMA analysis, profiles were clustered by calculating the log ratios of adjacent time-points, and the resulting matrix digitized by choosing a nominal threshold, 't'. Ratios were replaced by 1 if ratio > t, -1 if < -t and 0 otherwise. Profiles with the same digital pattern were assigned to the same cluster.
The output was tabulated and stored in a MySQL database and a custom web-interface was built to query the data using perl-cgi. The "Human Erythroblast Maturation (HEM) Database" is available at: https://cellline.molbiol.ox.ac.uk/eryth/index.html.

**QPCR**

cDNA was synthesized from 100ng RNA taken from individual samples using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, UK) following DNAse treatment. Semi-quantitative PCR (QPCR) was performed using TaqMan Universal PCR Master Mix (Applied Biosystems, UK) and intron-spanning primers and probes from the Roche Universal Probe Library (UPL, Table S1). Novel transcripts identified by microarray were selected for QPCR if an efficient UPL assay was available. Expression data for each transcript was normalized to that for the reference gene Ribosomal Protein L13A (RPL13A)\(^{26}\) and then quantified in Pro-Es, Int-Es and Late-Es relative to the expression level in CFU-Es, after running dilution standards for every assay to ensure the ratio of the differences in cycle thresholds \(2^{-\Delta\Delta CT}\) was equivalent to the fold change calculated from the standard curve.
Results

Sorted erythroblast populations exhibit expected cell surface marker expression and morphology

Erythroblasts were cultured from 83 Blood Transfusion Service (BTS) buffy coats and sorted, by CD36, CD71 and CD235a expression and size, into enriched CFU-E, Pro-E, Int-E and Late-E populations (Figure 1A). Typically 5x10^5–10^7 sorted cells were obtained in the final gates, with a typical RNA yield of 2µg from 10^6 cells. Post-sort analysis of two representative populations is shown in Figure S1.

May Grünwald-Giemsa-stained cytospins of cells illustrate the morphological changes within this culture as erythroblasts mature (Figure 1B). Maturation is associated with reduction in cell size, enhanced chromatin condensation and increased hemoglobinization. Ninety-seven sorts from a total of 54 donors gave 71 populations of suitable purity and morphology consistent with culture time. No non-erythroid cells were counted in these 71 populations, and fewer than 2% of the cells were of an earlier or later stage of maturation, as judged by inspection and counting the post-sort cytospins, and consistent with observations in a previous study which verified this method.14

Transcriptome Analysis

Forty-eight RNA samples from 38 donors in twelve pools of four were hybridized to twelve Affymetrix HGU133_plus_2.0 arrays, three each of CFU-E, Pro-E, Int-E, Late-E.

RMA analysis was first used to normalize the data.17 Expression levels of less than 100 in any replicate were considered to be below background. 8570 genes were expressed at a significant level at any stage of erythroblast maturation (Table S2), while a further 12094 genes represented on the HGU133_plus_2.0 arrays were absent from all stages studied. This equates to 24.5% PIDs on the array being present at any stage and is slightly more stringent than the percentage present called by the GCOS software before normalization, which ranged from 28-42%. The distribution of genes present at all combinations of stages is shown in Figure 2A. 8450 expressed genes were also present in previously published erythroid studies.4-13 The remaining 120 genes remained below an expression level of 500 at all stages. Figure 2B shows the distribution pattern of more highly expressed genes, including only those genes which achieved an expression value of 500 in all three pools at any one stage. Ingenuity Pathways Analysis was used to analyse the molecular and cellular functions enriched in these lists of genes present during erythroblast development (Figure S2 A and B).

The GCOS percentage present in the hybridizations of RNA from unsorted cells ranged from 38-45%. We normalized hybridization files from the unsorted RNA populations together with
those from the sorted populations. In the unsorted populations, 7352 genes were present in cultures at all time points, representing 75% of the total 9766 genes present at any time point. In sorted erythroblasts, 55% of the genes present at any stage were present at all stages, when normalized with unsorted cell data. This indicates that sorting the cells to eliminate non-erythroid cells and work with defined erythroblast populations considerably improves detection of differentially expressed transcripts.

Validation of erythroid gene expression by QPCR

QPCR was used to confirm that array signal intensity reflected the trend in gene expression for a selection of highly differentially expressed genes and genes expressed at a consistently high level (Figures 3 and S3).

For most of the differentially expressed genes analyzed, expression ratios were confirmed by QPCR of individual samples to be even greater than those observed using microarray hybridization of pooled samples, establishing our dataset as a useful resource. Many of the genes which appeared to be expressed at a consistently high level from the microarray data were shown by QPCR to be more differentially expressed. Previous investigators have also reported greater fold changes detected using QPCR compared with microarray for ALAS, GATA1 and KLF1 (Figure 3A) during erythropoiesis. Signiﬁcant discrepancies included HBB, which appeared to be highly expressed at all maturation stages following microarray hybridization, but QPCR conﬁrmed differential HBB expression as expected (Figure 3A). A large difference in HBB fold changes detected by microarray compared with QPCR during erythropoiesis has been previously reported, and it is commonly observed that QPCR is a more sensitive indicator of expression fold changes; the abundance of globin mRNA is well documented to interfere with the ability of microarray hybridizations to accurately detect signiﬁcant biomarkers in whole blood.

Differentially expressed genes: LIMMA

Following LIMMA analysis, pair-wise comparisons of gene expression between stages were considered significant if the B value was at least 2.945 (representing a 95% probability that the gene was differentially expressed) and expression levels in all three replicates at any stage were above 100. The pattern of B values for comparisons of expression of a PID between any pair of stages (CFU-E to Pro-E, CFU-E to Int-E, CFU-E to Late-E, Pro-E to Int-E, Pro-E to Late-E, Int-E to Late-E) was used as a basis of clustering (Table S3). This lowest stringency list of 3679 differentially expressed genes (Differentially Expressed 1 (DE1), Table S3) included a large cluster of 1180 terminally down-regulated genes which were signiﬁcantly down-regulated from CFU-E to Late-E, Pro-E to Late-E and Int-E to Late-E but no other combinations of pairs of stages. About 1/3 of genes in DE1 (1174) were differentially expressed in previously published erythroid studies.
and a further 258 were differentially expressed in our re-analysis of existing data.\textsuperscript{9,11}(Tables 1 and S3). Another 1096 genes were differentially expressed in maturing erythroblasts cultured in the absence of stem cell factor (SCF) which promotes erythroid expansion and delays differentiation,\textsuperscript{28} leaving the final 1/3 (1151) genes which have not been previously observed to be differentially expressed in maturing erythroblasts.

DE1 was refined by increasing the minimum average expression threshold to 1000 and imposing a minimum fold change threshold of 10-fold between stages. Finally, PIDs were selected whose minimum and maximum average expression levels between any stage varied by at least 1000, so that the most stringent list (DE_Sig, Table S4) of significantly differentially regulated transcripts comprises 327 genes which have a minimum average expression level of 1000 for at least one stage, a minimum fold change of 10-fold between any pair of stages, and a minimum difference of 1000 between maximum and minimum average expression at any stage. This list of differentially regulated genes contains many erythroblast-specific genes,\textsuperscript{24} and all of these 327 genes or their homologues have been previously observed to be expressed in at least one erythroid expression dataset.\textsuperscript{4-13} 168 genes in DE_Sig were previously shown to be differentially expressed in maturing erythroid cells.\textsuperscript{4-13} This is the first report of significant differential expression during erythropoiesis of the remaining 159 genes. Our re-analysis shows that 43 of these were also differentially expressed in existing datasets of maturing erythroblasts,\textsuperscript{9,11} and a further 110 were regulated in the absence of SCF.\textsuperscript{12} The remaining 6 genes were differentially expressed in our dataset only. Table 2 lists the functions of some of the genes which have not previously been described as differentially expressed in primary erythroblasts. QPCR was used to confirm expression profiles of selected significantly differentially regulated transcripts (Figures 3B and S3).

The majority of PIDs in DE_Sig, the most stringent list, fall into two clusters. Ninety-eight genes are terminally up-regulated (DE_SigUp) and 92 genes are terminally down-regulated (DE_SigDown) (Figure 4A, B, C and D).

A two-tailed t-test of the most significantly differentially regulated genes comparing the microarray expression ratios observed from the well-characterized sorted cell populations with those obtained from unsorted cells shows the significant differences between all six pair-wise stage ratios (Figure 4E) for the 327 genes. Sorting the cells permitted us to use homogeneous erythroblast populations enabling detection of a wider dynamic range of expression fold-changes of transcripts throughout maturation.

**Differentially expressed genes: Matched Filtering**

Alternative ways of clustering the data using a matched filtering approach produced a list of 91 genes which are continuously up-regulated between pairs of successive stages of erythroblasts studied (Matched Filter-derived list of up-regulated genes, MFUp) (Table S5). Essentially this
process can be summarized in three steps: 1) the median response for each PID at each stage was calculated; 2) three ratios of median responses (i.e. Pro-E/CFU-E, Int-E/Pro-E, and Late-E/Int-E) were calculated for each PID; 3) a list of genes was generated from the data by keeping the top 20% of the PIDs up-regulated between the CFU-Es and Pro-Es, then selecting the 20% of these PIDs which were most up-regulated between Pro-Es and Int-Es, then selecting the 20% of these PIDs which were most up-regulated between Int-Es and Late-Es.

Fifty-two of these genes have been previously reported as differentially expressed in maturing erythroid expression studies, with a further 23 detected by our re-analysis of existing datasets, 8 of them regulated only in the absence of SCF. Matched filter analysis identified two regulated genes not expressed in previous erythroid expression studies: gasdermin B (GSDMB), which was present only in Late-E, and zinc finger protein 805 (ZNF805). An additional 13 genes in MFUp were not previously differentially expressed. Table 2 lists the functions of some of the genes which have not previously been described as differentially expressed in primary erythroblasts, and Figure 3C shows QPCR confirmation of expression profiles of selected genes in MFUp.

**Consistently highly expressed transcripts**

A list of genes which are Consistently Highly Expressed (CHE) throughout erythroblast maturation was generated by considering those PIDs with a maximum expression value of at least 1000 for which all B values between the possible pair-wise comparisons were below -2.945 indicating a lower than 5% probability that the transcript was differentially regulated at any stage (CHE, Table S6). Any genes represented by divergent PIDs whose expression varied so much as to be included in both DE1 and CHE were removed from both lists. The final list of significant and continuously expressed transcripts contained 185 genes. Initial QPCR of splicing factor arginine-serine rich (SFRS2) and myosin light chain 12B (MYL12B) (Figure 3D) indicates that although neither of these transcripts have appeared to be differentially expressed during erythropoiesis in previously published datasets, their expression may vary more throughout erythroblast maturation than is indicated by the microarray data. This emphasises the need for verification of microarray data by alternative means. Expression of ATP synthase, H⁺ transporting, mitochondrial F0 complex, subunit G (ATP5L), eukaryotic translation elongation factor 1 alpha 1 (EEF1A1) and ribosomal protein S19 (RPS19) was confirmed by QPCR to be at a consistently high level, changing by less than 2-fold (Figure 3D). Therefore this set of genes (CHE) is a valid starting point for studying continuously expressed genes with significant expression levels throughout maturation. Our CHE list may be further refined by removing the 44 genes which were differentially regulated in previous studies, although 28 of these were differentially expressed only in the absence of SCF.
As with the DE_Sig list, many erythroblast-enriched genes\textsuperscript{24} are included in the CHE transcripts. All of these 185 genes or their homologues have been reported in previous erythroid expression datasets.\textsuperscript{4-13}

**Ontology Analysis of Significant Gene Lists**

Ingenuity Pathways Analysis and DAVID\textsuperscript{23} were used to analyse the molecular and cellular functions enriched in these gene lists. Genes affecting cell cycle were significantly represented in all three categories, genes controlling apoptosis and gene expression were significantly represented in terminally up-regulated genes, and genes controlling cellular assembly and organisation were significantly represented in terminally down-regulated genes (Figure S4). Protein synthesis was significantly associated with transcripts in both DE_SigUp and CHE gene lists but not with DE_SigDown, in keeping with the role of the red cell as a hemoglobin factory. Similarly, genes associated with RNA post-transcriptional modification were significantly represented in the CHE and DE_SigDown gene lists but not with DE_SigUp.

**Transcription Factor Binding Site analysis**

TFBS enriched in the lists CHE, DE_SigUp, DE_SigDown and MFUp were analyzed using DiRE, Pscan and CONFAC\textsuperscript{20-22} software (Figure 5, Table 3). All three packages query the TRANSFAC TFBS database. Pscan also queries the JASPAR database.

The CHE and DE_SigUp gene lists were both significantly enriched for binding sites of the CREB/G-box-like subclass of bZIP transcription factors (Figure 5, Table 3), and CHE was also significantly enriched for ETS class TFBS. Both classes are highly conserved and mediate diverse effects through many partners.\textsuperscript{29,30} As might be expected for genes that are switched off as cells become more specialized, fewer TFBS were found to be significantly enriched in the DE_SigDown list.
Discussion

Erythroid maturation is a gradual, continuous process, erythroblasts at different stages of maturation simultaneously co-existing in the bone marrow. CFU-E, Pro-E, Int-E and Late-E stages represent snapshots of the maturing erythroblast providing a tool to study gene expression during erythropoiesis. We have demonstrated that cells sorted with defined phenotypes express not only significantly fewer genes than unsorted cells, but also show significantly more differences of greater magnitude in the expression levels of particular genes between stages than unsorted cells. We have also identified many differentially expressed genes whose regulation has not been previously described in maturing primary erythroblasts. Finally, using transcription factor binding site analysis we have identified potential transcription factors that may regulate gene expression during terminal erythropoiesis.

The cRNA hybridized to the Affymetrix arrays was a pool of preparations from four morphologically and phenotypically homogeneous cell populations possessing a restricted range of expression of the erythroid markers CD36, CD71 and CD235a, therefore representing enriched erythroblast populations with the lowest possible level of contamination from other hematopoietic lineages.

Hybridizing pools of RNA from such rigorously defined populations clearly demonstrates significant increases in magnitude of expression ratios of differentially-expressed transcripts in RNA from sorted erythroblasts compared with that from unsorted cells from the two-stage cultures (Figure 4E). Removal of contaminating non-erythroid cells and enrichment of erythroblasts at a similar stage of maturation enables improved detection of expression ratios over a broader dynamic range. Re-analysis of other datasets from studies with a different emphasis\(^9\,12\) indicates that a smaller proportion of expressed genes were differentially expressed in non-sorted erythroblasts (Table 1), with the exception of erythroblasts cultured in the absence of SCF\(^12\) which would be expected to exhibit enhanced differentiation since SCF promotes erythroid expansion and delays differentiation.\(^28\) Furthermore, the robustness and integrity of our data is increased by the pooling strategy we used when preparing cRNA,\(^31\) as confirmed by QPCR of individual samples.

Genes most highly expressed and up-regulated during erythroid maturation include ribosomal and mitochondrial genes, those involved in transcription and translation, and other known erythroid genes such as globins and blood group antigens.

In our sorted erythroblasts, EPO-dependent transcripts previously identified using a cDNA subtraction library,\(^32\) including histones, high mobility group factors and transcription factors involved in chromatin remodelling, and DNA recombination and repair proteins such as topoisomerases, are mostly highly-expressed.
Keller and colleagues\(^9\) cultured CD34+ cells in a serum-free culture system containing EPO and harvested the unsorted cells at six time points between 1 and 11 days in culture, representing differentiation from cells more immature than CFU-Es to hemoglobinized cells appearing slightly less mature than our pyknotic Late-Es. They identified 1500 differentially expressed genes, approximately 20% of which are in DE1, our least stringent list of 3678 differentially regulated transcripts, while 93 genes are in our most stringent differentially regulated list DE_Sig which contains 327 genes. Many of the discrepancies between these sets of differentially regulated genes may be due to the differences in maturity of the erythroblasts studied,\(^9\) or to specific culture conditions. Our dataset includes 1151 differentially regulated genes which were not differentially regulated during erythropoiesis in previous studies\(^4\)-\(^{13}\) plus a further 1096 which were differentially expressed only when erythroblasts were cultured in the absence of SCF, conditions which would enhance erythroid differentiation.\(^28\) This is probably because we defined populations by expression of CD36, CD71 and CD235a, resulting in a wider dynamic range of detectable fold-change ratios (Figure 4E) and a lower proportion of transcripts apparently present at all stages of maturation. The significance of expression profiles of specific genes will become apparent as more expression data and, in particular, functional studies become available.

Some of the significantly regulated genes we have identified, such as SMARCA4 which co-activates the gamma-globin gene, have known erythroid functions (Tables 2 and S8A). The erythropoietic role of others, known to have relevant functions such as inhibition of apoptosis\(^{33,34}\) or proliferation\(^{35}\) in other cell types, must be confirmed by functional studies. Most of the significantly differentially expressed genes identified in our dataset were terminally differentially regulated (Figure 4A-D), and therefore probably beyond the scope of previous studies except for the study omitting SCF from the culture medium.\(^{12}\) An example of a novel erythroid expression profile identified by our study is that of the inhibitor of apoptosis protein gene baculoviral IAP repeat-containing 3 (BIRC3), which blocks caspase-induced apoptosis by direct inhibition of caspases 3 and 7.\(^{36}\) Previously observed to be down-regulated during the earliest stages of erythropoiesis and then remain expressed at a low level,\(^9\) BIRC3 is in DE_SigUp, significantly up-regulated between Int-Es and Late-Es and uniquely and substantially expressed in Late-Es. This expression pattern was confirmed by QPCR (Figure 3A). This difference is perhaps because the induction we observe is in homogeneous erythroblasts more mature than those in previous studies.\(^4\)-\(^{13}\)

RFXAP, significantly up-regulated in our Int-Es as confirmed by QPCR (Figure 3B), was not differentially regulated during erythropoiesis in previous studies,\(^9\)-\(^{11,13}\) although our retrospective analysis of previous data shows that it was up-regulated during maturation of erythroblasts cultured in the absence of SCF.\(^{12}\) The transcription factor RFXAP, in association with NF-Y and CREB, is involved in MHC class II activation.\(^{37}\) Binding sites for both CREB and NF-Y are enriched
in promoters of genes in DE_Sig and also of genes expressed at a consistently high level in our maturing erythroblasts (CHE, Figure 5, Table 4). RFXAP regulation during erythroblast maturation suggests it may have a novel role in erythropoiesis. Activating transcription factor 3 (ATF3), a CREB transcription factor, is also significantly up-regulated, and was one of only 20 genes in DE_Sig which is consistently significantly up-regulated between all pairs of stages studied except for CFU-E to Pro-E. ATF3 is involved in apoptosis, growth, commitment, proliferation, cell cycle progression and morphology and represses the erythroid-derived 2–related factor 2 (Nrf2)-regulated stress response pathway.38 Erythropoiesis-related functions of selected transcription factors highlighted in our TFBS analysis of CHE and DE_SigUp, many of which are consistent with enriched molecular function analysis results (Figure S4), are listed in Table 3B.

The 8570 genes present in the sorted erythroid cells contain 496 genes whose expression has been reported to be specific to non-erythroid hematopoietic lineages.24 This HaemAtlas study compared erythroblasts at a stage between our Pro-E and Int-E stages39 with cells from other hematopoietic lineages, and lists lineage-specific transcripts. However, in that study, the definition of lineage-specific transcripts required at least two-fold up-regulation in the respective lineage.24 Therefore, some erythroid transcripts may have been present at a higher level in non-erythroid lineages or absent at the specific time-point studied. Erythroid expression of many of these 496 transcripts we have detected has also been reported elsewhere including the HaemAtlas study itself4-13,32,40 (Table S7). None of the markers used to select the non-erythroid lineages24 are present in our dataset, with the exception of Integrin α2B (ITGA2B, the platelet antigen CD41), which reached an average expression level of 350 in CFU-Es, and was also present in erythroblasts in the HaemAtlas study.24

Of the 127 genes present only in CFU-Es with an expression cut-off of 100 (Figure 2A), only six exceeded a minimum expression level between the three CFU-E pools of 200. None reached an expression level of 500. Genes in this category are likely to be those which are more highly expressed during earlier developmental stages, being enriched for Hematological System Development and Function (24 associated transcripts, p=1.55 x 10^{-5}) (Figure 2C). These include Ets variant 6 (ETV6), which has a role in expanding erythroid precursors and accumulating hemoglobin,41 and hematopoietically expressed homeobox (HHEX) which regulates proliferation and differentiation of hemangioblasts and endothelial cells during ES cell differentiation.42 The mouse homologue, Hhex, was rapidly down-regulated when erythroid differentiation was induced by GATA1 in G1ER cells.7 The most significant molecular and cellular function associated with this CFU-E-specific gene list is cellular growth and proliferation (29 transcripts associated with this function, p=1.55 x 10^{-5}), which is entirely consistent with the role of HHEX and ETV6 in CFU-E maturation. Nineteen genes present only in CFU-Es, including ETV6, have previously been
reported to be down-regulated during early erythropoiesis, indicating that they were up-regulated at earlier stages of erythroid development.

Of the 511 genes present only in Late-Es with an expression cut-off of 100 (Figure 2A), seven exceed a minimum expression level of 1000. Four of these genes are associated with apoptosis: tribbles homolog 3 (TRIB3), heme oxygenase 1 (HMOX1), BIRC3 and B-cell CLL/lymphoma 6 (BCL6). The molecular and cellular function Cell Death is significantly enriched in genes specifically expressed only in Late-Es (78 associated transcripts, p=3.91 x 10^{-4}) (Figure S2A). Genes uniquely expressed in Late-Es are associated with a role in gene expression (77 associated transcripts, p=5.09 x 10^{-5}) (Figure S2A).

The 1144 genes exceeding a minimum expression level of 500 at all stages (Figure 2B) are significantly enriched for the molecular and cellular functions “Protein synthesis” (127 associated transcripts, p=4.83 x 10^{-34}), “RNA post-transcriptional modification” (83 associated transcripts, 5.58 x 10^{-17}), “Post-translational modification” (84 associated transcripts, p=1.1 x 10^{-8}) and “Protein folding” (21 associated transcripts, p=1.1 x 10^{-8}). This is compatible with terminal differentiation and the high level of globin synthesis required.

Of 123 genes within 60 kb of SNPs found to be associated with blood cell traits, 80 genes were expressed at any stage throughout erythropoiesis, 20 were differentially expressed at the least stringent level in DE1, and three at the most stringent level in DE_Sig: membrane-associated ring finger (C3HC4) 8 (MARCH8), myeloblastosis oncogene (MYB), and transmembrane and coiled-coil domain family 2 (TMCC2). The CHE gene list contained sialomucin CD164, TFRC, and protein tyrosine phosphatase, non-receptor type 11 (PTPN11).

We have begun to explore different approaches of identifying co-regulated genes, for example, consistently up-regulated genes. In the broad framework of signal processing, it may be possible to design a matched filter to describe such a trend and to capture the effects of the relevant biochemical processes. In reality, however, it is much more complicated, in that not all up-regulated genes will experience equal up-regulation. Comparison of the successively up-regulated genes identified by matched filtering (MFUp) demonstrated the potential danger of missing important genes if the wrong type of analysis is used. While the matched filtering approach identified 47 consistently up-regulated genes in our data set which were also identified as significantly differentially regulated when selected on the basis of their B values (DE_Sig), and 98 genes whose erythroid expression has been previously reported, several new differentially expressed genes were also identified by this approach, with the expression profiles of BRWD3, CTS12, DNAJB4, GSDMB, LIN7B, RRAD, RFFL and ZNF805 confirmed by QPCR (Figure 3C). These eight genes are expressed at modest levels compared to many of those already described in erythroid cells, and future functional studies will be necessary to determine their role in erythropoiesis. However, many of them have known roles in autophagy, apoptosis and
transcriptional regulation\textsuperscript{45-52} (Tables 2 and S8A) which may be relevant to the maturing erythroblast. Importantly, this approach enabled selection of genes based on their CFU-E to Pro-E expression ratios; to do this using the LIMMA approach would have required lowering the selection stringency of B and/or p-values. Future studies will consider that not all up-regulated genes experience equal up-regulation, and extend the design of adaptive matched filters\textsuperscript{53} to describe a trend. Beyond the novelty of matched filtering, we intend to continue to explore such new approaches (like adaptive matched filtering) that have general application to microarray biological datasets.

Microarray technology is less sensitive than QPCR at quantifying expression levels, with a more restricted dynamic range. Signal intensity may be affected by location and hybridization efficiency of probes, yet arrays have the obvious advantages of large-scale analysis of gene expression. Deep sequencing will supersede microarray hybridizations as the optimum method for large scale transcriptome analysis. However, our dataset represents a novel and valuable resource to explore gene expression changes during erythroid maturation, and to compare with expression data from the diseased state.

We studied enriched erythroblast populations with the lowest possible level of contamination from other hematopoietic lineages to show that cells sorted with defined phenotypes express significantly fewer genes than unsorted cells, and show significantly more differences of greater magnitude in the expression levels of particular genes between stages. We have also identified many differentially expressed genes previously unregulated in erythroid expression studies with unknown roles in erythropoiesis and highlighted potential transcription factors that may regulate gene expression during terminal erythropoiesis. The HEM database is freely available at https://cellline.molbiol.ox.ac.uk/eryth/index.html.
Acknowledgements

We are grateful to Chris Fisher, Jim Hughes, Abigail Lamikanra, Nick Watkins and David Weatherall for helpful discussion, Tariq Enver for support, Helen Cattan and Emanuele Marchi for bioinformatics advice, Jackie Sloane-Stanley, Sue Butler and Jill Brown for technical assistance, and Giulio Pavesi and Carlos Moreno for email correspondence regarding Pscan and CONFAC respectively.

This report presents independent research funded in part by NHS Blood and Transplant, National Institute for Health Research (NIHR) Biomedical Research Centre Programme and by NIHR under its Programme Grants for Applied Research scheme (RP-PG-0310-1004). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Authorship

ATMC and KJHR designed and performed research, analyzed data and wrote the paper. AA performed research and analyzed data. KC and CW performed research. SS, NG, & ST analyzed data. VB & SJMcG analyzed data and contributed to writing the paper. AKN analyzed data and wrote the paper. WGW and DRH designed research. DJR wrote the paper. The authors have no conflicts of interest.
References


Table 1

We analyzed other datasets\textsuperscript{9-12} deposited with NCBIGEO containing expression measurements from the indicated number of replicates and stages of erythroblasts cultured as summarized. RMA (where the CEL files are available) and LIMMA were used to determine the proportion of genes that were:

Differentially expressed (DE) at the lowest stringency equal to DE1 (B value of at least 2.945 (representing a 95% probability that the gene was differentially expressed) and expression levels in all replicates at any stage above 100)

Significantly differentially expressed (SIG DE) at the highest stringency equal to DE_Sig (minimum average expression threshold of 1000, minimum fold change threshold of 10-fold, only PIDs with a difference of at least 1000 between minimum and maximum average expression were selected)

<table>
<thead>
<tr>
<th>Study</th>
<th>NCBIGEO accession no.</th>
<th>CEL files deposited?</th>
<th>No. stages</th>
<th>No. Replicates</th>
<th>Pooled or single samples hybridized?</th>
<th>EB source &amp; culture conditions</th>
<th>FACSSorted</th>
<th>% Genes DE</th>
<th>% Genes SIG DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keller et al 2006</td>
<td>GSE4655\textsuperscript{9}</td>
<td>No</td>
<td>6</td>
<td>3</td>
<td>Some pooled samples</td>
<td>CD34+ progenitors in a novel serum-free system</td>
<td>No</td>
<td>12</td>
<td>3.8</td>
</tr>
<tr>
<td>Singleton et al 2008</td>
<td>GSE10584\textsuperscript{11}</td>
<td>Yes</td>
<td>3</td>
<td>4</td>
<td>Single samples</td>
<td>CD34+ progenitors</td>
<td>No</td>
<td>8</td>
<td>0.1</td>
</tr>
<tr>
<td>Peller et al 2009</td>
<td>GSE12407\textsuperscript{10}</td>
<td>Yes</td>
<td>3</td>
<td>2</td>
<td>Single samples</td>
<td>Fibachs</td>
<td>No</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Sripichai et al 2009</td>
<td>GSE7874\textsuperscript{12}</td>
<td>Yes</td>
<td>2</td>
<td>5</td>
<td>Single samples</td>
<td>CD34+ progenitors cultured for 7 and 14 days +EPO/-SCF</td>
<td>No</td>
<td>47</td>
<td>0.3</td>
</tr>
<tr>
<td>Current Study</td>
<td>GSE22552</td>
<td>Yes</td>
<td>4</td>
<td>3</td>
<td>Pooled samples</td>
<td>Fibachs</td>
<td>Yes</td>
<td>43</td>
<td>3.8</td>
</tr>
</tbody>
</table>
Table 2

Selected genes in DE_SigUp or MFUp which have not been previously described as differentially regulated in erythroblasts. The largest significant fold change between stages of our maturing erythroblasts is indicated (C, CFU-E; P, Pro-E; I, Int-E; L, Late-E) with a brief summary of the function and whether it was differentially expressed in any of the re-analyzed datasets listed in Table 1, or any of the previous erythroid cell line studies. References for gene function are given in Table S8B.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Symbol</th>
<th>List</th>
<th>Largest significant FC between erythroid stages</th>
<th>Function</th>
<th>DE in a re-analyzed dataset?</th>
<th>DE in a previous cell line study?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbonic anhydrase 3</td>
<td>CA3</td>
<td>DE_Sig</td>
<td>177.7x C-I</td>
<td>Muscle-specific, catalyses reversible hydration of CO₂, inhibits apoptosis</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Calbindin 1</td>
<td>CALB1</td>
<td>DE_Sig</td>
<td>173.1x I-L</td>
<td>Inhibits apoptosis in osteoblasts by binding to and inhibiting Caspase 3</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>FK506 binding protein 1B, 12.6 kDa</td>
<td>FKBPIB</td>
<td>DE_Sig</td>
<td>13.1x C-L</td>
<td>Inhibits the mTOR proliferation pathway</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Intalectin 1</td>
<td>ITLN1</td>
<td>DE_Sig</td>
<td>31.3 C-L</td>
<td>Binds cell surface galactofuranidase</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Multiple EGF-like-domains 9</td>
<td>MEGF9</td>
<td>DE_Sig</td>
<td>18.1x C-L</td>
<td>Signalling molecule, regulated during development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-myc downstream regulated 1</td>
<td>NDRG1</td>
<td>DE_Sig</td>
<td>10.5x C-I</td>
<td>Involved in hypoxia response, cell differentiation and cell growth arrest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTEN-induced putative kinase 1</td>
<td>PINK1</td>
<td>DE_Sig</td>
<td>11.5x C-L</td>
<td>Implicated in mitochondrial quality control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulatory factor X-associated protein</td>
<td>RFXAP</td>
<td>DE_Sig</td>
<td>-32.9x I-L</td>
<td>In association with NFY and Creb, Involved in MHC class II activation</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>S100 calcium binding protein A4</td>
<td>S100A4</td>
<td>DE_Sig</td>
<td>-24.5x P-L</td>
<td>Over-expression promotes metastasis, activates NF-KappaB, Functions as a Ca²⁺ activated molecular switch, regulates macrophage recruitment and chemotaxis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SWI/SNF related, matrix associated, actin</td>
<td>SMARCA4</td>
<td>DE_Sig</td>
<td>-11.0x C-L</td>
<td>Co-activates gamma-globin gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Name</td>
<td>Gene Symbol</td>
<td>Expression</td>
<td>Function</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>------------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependent regulator of chromatin, subfamily a, member 4</td>
<td>TMEM140</td>
<td>DE_Sig 26.6x C-L</td>
<td>Potential role in supporting HSC activity of a stromal cell line</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zinc finger protein 451</td>
<td>ZNF451</td>
<td>DE_Sig 11.5x C-L</td>
<td>ZNF451 is a novel PML body- and SUMO-associated transcriptional co-regulator</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histone cluster 1, H2bd</td>
<td>HIST1H2BD</td>
<td>DE_Sig &amp; MFUp 26.0x C-L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bromodomain and WD repeat domain containing 3</td>
<td>BRWD3</td>
<td>MFUp 6.7x C-L</td>
<td>May have a role in JAK/STAT signalling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cathepsin L2</td>
<td>CTSL2</td>
<td>MFUp 7.1x C-L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DnaJ (Hsp40) homolog, subfamily B, member 4</td>
<td>DNAJB4</td>
<td>MFUp 15.3 C-L</td>
<td>A tumor suppressor activated by YY1 and AP1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gasdermin B</td>
<td>GSDMB</td>
<td>MFUp 8.8 C-L</td>
<td>Involved in regulation of apoptosis in the gastrointestinal epithelium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Importin 11</td>
<td>IPO11</td>
<td>MFUp 12.1x C-L</td>
<td>Increases the intensity of pleiotrophic cytokine responses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lin-7 homolog B (C. elegans)</td>
<td>LIN7B</td>
<td>MFUp 19.3 C-L</td>
<td>Involved in Ras signalling; effects include cell growth, differentiation and survival</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear paraspeckle assembly transcript 1 (non-protein coding)</td>
<td>NEAT1</td>
<td>MFUp 28.7x C-L</td>
<td>RNA component of paraspeckles, ribonucleoprotein structures involved in regulating the expression of certain genes in differentiated cells by nuclear retention of RNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retinoic acid receptor responder (tazarotene induced) 3</td>
<td>RARRES3</td>
<td>MFUp 5.2 C-L</td>
<td>Retinoid-inducible growth regulator, induces apoptosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rififylin</td>
<td>RFFL</td>
<td>MFUp 7.7 C-L</td>
<td>Inhibits recycling from the endocytic recycling compartment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ras-related associated with diabetes</td>
<td>RRAD</td>
<td>MFUp 12.5 P-L</td>
<td>Involved in Ras signalling; effects include cell growth, differentiation and survival</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lysine (K)-specific demethylase 6A</td>
<td>KDM6A</td>
<td>MFUp 7.8 C-L</td>
<td>Involved in chromatin remodelling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zinc finger protein 805</td>
<td>ZNF805</td>
<td>MFUp 6.2 C-L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3
Summary of transcription factor binding site analysis

Table 3A
Two transcription factor classes significantly represented in genes in CHE and DE_SigUp. Bonferroni p-value and Information content returned by Pscan are given.

<table>
<thead>
<tr>
<th>TFBS</th>
<th>Bonferroni p-value in CHE</th>
<th>Bonferroni p-value in DE_SigUp</th>
<th>Information content</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREB/G-box-like bZIP</td>
<td>5.85E-08</td>
<td>8.00E-03</td>
<td>10.79</td>
</tr>
<tr>
<td>ETS Family</td>
<td>1.31E-08</td>
<td>ns</td>
<td>8.99</td>
</tr>
</tbody>
</table>

Table 3B
Selected transcription factors significantly represented in genes in CHE and DE_SigUp with examples of functions

<table>
<thead>
<tr>
<th>TF</th>
<th>TFBS enriched in</th>
<th>Erythroid-related functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREB</td>
<td>CHE, DE_SigUp</td>
<td>Key mediator of myeloid differentiation &amp; proliferation.</td>
</tr>
<tr>
<td>AP2/ERF</td>
<td>CHE</td>
<td>Transcriptional repressor, induces erythroid differentiation in K562 and HEL cells.</td>
</tr>
<tr>
<td>E2F</td>
<td>CHE, DE_SigUp</td>
<td>E2F plays critical roles in cell cycle progression by regulating the expression of genes involved in nucleotide synthesis, DNA replication, and cell cycle control as demonstrated in mouse models. E2F has recently been shown to form a complex with GATA1, FOG1, and retinoblastoma protein (pRb), thus stalling cell proliferation and steering erythroid precursors towards terminal differentiation.</td>
</tr>
<tr>
<td>NFY(A) (CAAT-binding protein)</td>
<td>CHE, DE_SigUp</td>
<td>Binds human beta-globin promoter in K562 cells, stabilising a complex formed with MZF1 and GATA2, appears important in maintaining a chromatin conformation favourable to beta-globin transcription.</td>
</tr>
<tr>
<td>SP1</td>
<td>CHE, DE_SigUp</td>
<td>Ubiquitously expressed. Acts synergistically with FOXM1, GATA1, OCT &amp; AP2 which are also highlighted in DiRE analysis of DE_SigUp.</td>
</tr>
<tr>
<td>TATA/TBP</td>
<td>DE_SigUp</td>
<td>TBP is essential for long-range activation of beta-globin transcription in response to the HS2 enhancer located 10kb from the promoter.</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1 Preparation of enriched populations of CFU-Es, Pro-Es, Int-Es and Late-Es.

A. CD36, CD71 (transferrin receptor 1) and CD235a expression determined by flow cytometry of PBMC-derived erythroblasts. Cells were stained for 15’ at 4°C in 2% BSA in HBSS and washed twice. CFU-Es are CD36+, CD71+ and CD235a-low. CD235a expression increases until the intermediate erythroblast stage, when it remains high and CD71 expression begins to decrease. Stained cells were sorted on a MoFlo cell sorter (Beckman Coulter) to obtain erythroblast populations enriched for the appropriate stage of development. Our cells were tightly gated to ensure discrete populations of cells similar in maturity and lineage. Sort gates were defined by first setting a gate on the fluorescent expression profile of the population of interest, either CD71 vs CD36 (i) and/or CD71 vs CD235a (ii-v); these gates were then applied to the Forward Scatter vs Side Scatter dot plot and a 2nd gate was applied to exclude debris and also to define a population of cells showing similar size. The mean percentages of sorted cells collected in the total erythroid gates were 14%, 24%, 20% and 50% for cells sorted on days 4, 5-6, 8-11 and 13-15, respectively. Sorted CFU-Es, Pro-Es, Int-Es and Late-Es represented an average of 28%, 29%, 37% and 45% of erythroid cells present, respectively.

B. Cytospins of cells stained with May Grünwald-Giemsa are shown to illustrate the morphological changes as cultured erythroblasts mature from large CFU-Es with a large, smooth nucleus at day 4, through the Pro-E and basophilic Int-E stages as the chromatin becomes more condensed, accompanied by a reduction in both cell size and the nucleus:cell size ratio, to the small Late-Es when the cytoplasm turns from blue to orange reflecting the production of hemoglobin and the nuclei become pyknotic. From all hybridized sorts, we have counted at least 250 cells per sample and find an average of 99.3% purity for our populations, with the contaminants being erythroid cells of slightly increased or decreased maturity rather than non-erythroid cells.

Representative plots and cytospins viewed at 20x magnification are shown.

Figure 2

Four-way Venn diagram illustrating the overlap of genes expressed in all pools at different stages of erythroblast maturation. Expression cut-offs are selected with the caveat that expression measurements on array platforms are relative and prone to influence by physical hybridization factors.

Numbers of transcripts are shown which exceeded an expression level of 100 in all three triplicates at each stage. 12094 genes represented on the HGU133_plus_2.0 arrays were absent from all stages studied.
A. Numbers of transcripts that exceeded an expression level of 500 in all three triplicates at the combinations of stages indicated.

Four thousand six hundred and forty genes are common to all stages, while other genes show a more stage-specific expression pattern, for example 511 are only present at the Late-E stage (red). When the expression cut-off is increased to 500, the number of transcripts meeting that criteria at all stages drops to 1144, and 396 genes exceed an expression level of 500 at the Late-E stage only.

Figure 3
QPCR confirmation of transcript fold changes from microarray expression profiling. Transcripts were selected for QPCR based on the availability of an efficient intron-spanning QPCR assay. For each gene, the fold change is given between CFU-E and Pro-E (C-P, green bars), Int-E (C-I, blue bars) and Late-E (C-L, pink bars). Solid bars on the left-hand side of each graph represent microarray data, hatched bars on the right represent QPCR data in triplicate from at least three biological replicates of individual samples, including some additional samples not included in the pools on the Affymetrix arrays.

A. Previously described erythroid genes;
B. Differentially regulated transcripts from DE_Sig;
C. Differentially regulated transcripts identified by Matched Filter analysis (MFUp);
D. Transcripts which are continuously expressed at a high level throughout maturation (CHE);

Figure 4
A. Key to the five-way Venn diagram in Figure 4B of the 384 most significantly differentially expressed genes (DE_Sig) showing the overlap of fold change combinations. The three circles represent transcripts differentially expressed between Late-Es (L) and one of CFU-E (C, green), Pro-E (P, blue) or Int-E (I, red). The wide magenta arc represents genes differentially regulated between Pro-Es and Int-Es. The narrower yellow arc encompassing most of the perimeter of the wider arc represents genes differentially regulated between CFU-Es and Int-Es. The colours merge where the shapes intersect each other.

B. Five-way Venn diagram of the 384 most significantly differentially expressed transcripts showing the overlap of fold change combinations. See Figure 4A for the key. The number of transcripts which are up-regulated during maturation is shown in white numbers with a dark shadow, while the number of transcripts that are down-regulated during maturation are shown in black numbers with a white shadow. Most genes are either terminally up-regulated or terminally down-regulated. For example, 98 transcripts are up-regulated between CFU-Es to Late-Es, Pro-Es to Late-Es to Int-Es to Late-Es, and 92 transcripts are
down-regulated between CFU-Es to Late-Es, Pro-Es to Late-Es to Int-Es to Late-Es. One transcript is up-regulated between CFU-Es to Int-Es and Pro-Es to Int-Es, but not between any of these stages and Late-Es. Where no number is shown, there were no transcripts in that set.

C. Heat maps of expression levels of transcripts most significantly differentially expressed (DE_Sig) up-regulated during erythroid maturation. Each column represents a different stage of erythroblast maturation, and each row represents a different PID. Red, white and blue left-hand panels indicate log (base 2) of expression ratios between the stage indicated in the column heading and CFU-Es during erythroid maturation. White represents a 1:1 ratio, so the CFU-E column is always white, red indicates up-regulation (log fold change of 3 indicating an 8-fold increase during maturation), and blue indicates down-regulation (log fold change of -3 indicating an 8-fold decrease during maturation). Red and blue right-hand panels indicate linear expression levels. Blue indicates no or low expression, and red indicates high expression (above 1000). Expression dynamics are indicated.

D. Heat maps of expression levels of transcripts most significantly differentially expressed (DE_Sig) down-regulated during erythroid maturation. Each column represents a different stage of erythroblast maturation, and each row represents a different PID. The red, white and blue left-hand panel indicates log (base 2) of expression ratios between the stage indicated in the column heading and CFU-Es during erythroid maturation. White represents a 1:1 ratio, so the CFU-E column is always white, red indicates up-regulation (log fold change of 3 indicating an 8-fold increase during maturation), and blue indicates down-regulation (log fold change of -3 indicating an 8-fold decrease during maturation). The red and blue right-hand panel indicates linear expression levels. Blue indicates no or low expression, and red indicates high expression (above 1000). Expression dynamics are indicated.

E. Significant increase in the range and magnitude of fold changes of expression of the 384 most differentially-expressed transcripts (DE_Sig) from CFU-E/Pro-E (C-P), CFU-E:/Int-E (C-I), CFU-E/Late-E (C-L), Pro-E/Int-E (P-I), Pro-E/Late-E (P-L) or Int-E/Late-E (I-L) in RNA from sorted erythroblasts (S) compared with that from unsorted cells (U). A two-tailed paired t-test was performed. Between each pair of stages compared, the detected fold changes were distributed over a narrower range of values in the unsorted erythroblasts compared with their sorted counterparts, indicating a dampening of the variation detected between different stages of erythroblasts in the unsorted samples.
Figure 5

A. Transcription Factor Analysis of genes that are consistently highly expressed during erythroid maturation (CHE). Significant TFBS from Pscan output querying the JASPAR and TRANSFAC databases are shown. Green bars indicate bZIP TFBS, red bars indicate ETS Factor TFBS, solid bars indicate the TFBS was detected as significant by more than one of the three programmes Pscan, DiRE and CONFAC, hashed bars indicate the TFBS was detected as significant by Pscan only.

B. Transcription Factor Analysis of genes that are significantly up-regulated during erythroid maturation (DE_SigUp).

   i. Significant TFBS from Pscan output querying the JASPAR and TRANSFAC databases;

   ii. Significant TFBS from DiRE output (only TFBS which are also present in Pscan or CONFAC output or in bZIP factor family returned by Pscan search against JASPARFam database are shown);

   iii. Significant TFBS from CONFAC output returned with a mean difference of the number of TFBS for each promoter between sample and control greater than 0.5 (only TFBS which are also present in Pscan or DiRE outputs or in bZIP or ETS factor families returned by Pscan search against JASPARFam database are shown).

Green bars indicate bZIP TFBS, solid bars indicate the TFBS was detected as significant by more than one of the three programmes Pscan, DiRE and CONFAC, hashed bars indicate the TFBS was detected as significant by a single programme only.
Figure 1A

i CD36:CD71 (d4-11)

ii CFU-E d4

iii Pro-E_d6

iv Int-E_d11

v Late-E_d14

Figure 1B

i CFU-E d4

ii Pro-E_d6

iii Int-E_d11

iv Late-E d14
Figure 3B

CA3
CALB1
FKBP1B
HIST1H2BD
ITLN1
PINK1
RFXAP
SMARCA4
TMEM140
Figure 3C

BRWD3

CTSL2

DNAJB4

GSDMB

LIN7B

RFFL

RRAD

ZNF805

FC from CFU-E
Figure 3D

- ATP5L
- EEF1A1
- MYL12
- RPS19
- SFRS2
Figure 4B
Two-tailed paired t-test of pairwise fold changes for the 384 most significantly differentially regulated PIDs, comparing expression ratios from sorted and unsorted populations.
Figure 5B

-Log Bonferroni p-value

-Log p-value

Importance
Global gene expression analysis of human erythroid progenitors