SUPPLEMENTAL MATERIAL

Developmental transcriptome analysis of human erythropoiesis

Lihong Shi, Yu-Hsuan Lin, M.C. Sierant, Kim-Chew Lim, Fan Zhu,

Shuaiying Cui, Yuanfang Guan, Maureen A. Sartor,

Osamu Tanabe and James Douglas Engel

SUPPLEMENTAL FIGURES

Supplemental Figure 1. *Ex vivo* erythroid differentiation of human CD34$^+$ hematopoietic progenitor cells was confirmed by flow-cytometric analysis. Dot plots show the expression of hematopoietic progenitor marker CD34, the thrombospondin receptor (CD36), transferrin receptor (CD71), glycophorin A (GPA) expression during *ex vivo* erythroid differentiation. Gates were defined using isotype-specific antibodies as controls. CD34$^+$ cells precipitously declined to less than 1% by day 14 (**top panel**) whereas cells double-positive for CD71 and GPA increased from 66% to 99% between days 4 and 14 (**bottom panel**). The fraction of differentiating cells expressing CD36 antigen transiently peaks between days 8 and 11 and begins to decline by day 14 (**top panel**).

Supplemental Figure 2. Validation of RNA-seq data.

(A) Reproducibility of RNA-seq assays between two biological replicates was assessed. The scatter plots represent the correlation of the RNA abundance between the data sets from two independent erythroid cell cultures prepared from CD34$^+$ cells that were subjected to differentiation conditions for 4, 8, 11, or 14 days. Binary logarithms ($\log_2$) of RNA abundances in FPKM (fragments per kilobase of exon per million fragments

mapped) are shown. The dot color corresponds to expression level differences between the two data sets: black dots differed by less than 1.5-fold, green dots differed by 1.5 to 2-fold, red dots differed by 2 to 4-fold, and blue dots differed by greater than 4-fold; “r” refers to Pearson’s correlation coefficient. The data indicate that there is close correlation between two biological replicates in their gene expression profiles, with r values greater than 0.97 on all days. (**B**) RNA quantification by RNA-seq was validated by reverse transcription and quantitative real-time PCR (qRT-PCR). The relative abundance of 37 genes, normalized to 4 days of *in vitro* culture, in the same RNA preparations was determined by both RNA-seq and qRT-PCR. Binary logarithms (log_2) of the relative expression levels are presented. The data reflect a very close correlation between the two data sets; “r” refers to Pearson’s correlation coefficient.

Supplemental Figure 3. A putative novel protein-coding gene potentially generated by gene duplication. (**A**) The top panel shows the location and expression of transcript TCONS_00040131. In this case, transcribed from an intergenic region with multiple exons at substantial expression levels during erythroid differentiation. The red arches indicate the exon junctions identified at day 4 of the differentiation culture. The lower panel shows the absence of this transcript in the USCS, RefSeq, or human EST databases. A protein BLAST search identified a similarity between its translation product and the FANCD2 (Fanconi anemia, complementation group D2) protein. These observations suggest that the novel transcript TCONS_00040131 is encoded by a previously unidentified protein-coding gene localized in an intergenic region. The y-axis represents the number of reads mapped to each genomic location, and ranges from 0 to 269 for all differentiation stages. (**B**) A nucleotide sequence alignment of the transcript TCONS_00040131 (top) and the FANCD2 mRNA (RefSeq accession number: NM_001018115.1, bottom) revealed 97% nt sequence identity in a 1,046-nt region

comprising 24% (nts 1,490 through 2,535) of the FANCD2 mRNA sequence. An amino acid sequence alignment of a translation product of TCONS_00040131 and the FANCD2 protein revealed 95% amino acid sequence identity over a 313-amino acid region comprising 22% (amino acids 507 through 819) of the FANCD2 protein (not shown). An overall comparison between TCONS_00040131 and the FANCD2 transcripts indicates that they are both located on chromosome 3 (separated by 1,769,054 bp) and are comprised of 10 *vs*. 44 exons, mature mRNAs of 1146 *vs*. 5134 nt and ORFs of 939 *vs*. 4353 amino acids, respectively.

Supplemental Figure 4. Validation of putative novel protein-coding transcripts. (**A**) Nine examples of the most probable novel protein-coding transcripts generated from intergenic regions are listed with data for each criterion applied to identify these transcripts. (**B**) The bar graph shows expression levels of these putative novel protein-coding transcripts at four differentiation stages on a binary logarithmic scale. (**C**) The existence of these transcripts in primary human erythroid cells were verified by RT-PCR (reverse transcription and PCR) using primers that span predicted exon junctions (**Supplemental Table 10**) using a pool of cDNAs from the four differentiation stages as PCR templates. Authenticity of PCR amplicons were confirmed by size determination by agarose-gel electrophoresis as shown here, and by Sanger sequencing of the amplicons with the primers used for PCR. Predicted amplicon sizes are indicated at the bottom of the gel picture.

Supplemental Figure 5. Functional analysis of 815 highly expressed transcripts. (**A**) Gene Ontology (GO) terms (33) which show statistically significant enrichment in the most highly expressed transcripts (expression levels > 128 FPKM) at each differentiation stage are plotting with their *P*-values, determined using a modified Fisher's exact test, on

a logarithmic (log_{10}) scale. ConceptGen (78) was used to identify over-represented GO terms. (**B**) Listed are the ten most significantly enriched GO terms in the 308 transcripts with constitutive high-level expression (expression levels > 128 FPKM at all four differentiation stages) during erythropoiesis plotting against *P*-value on a logarithmic (log_{10}) scale. (**C**) KEGG pathway analysis (80) indicates that the “ribosome” pathway is significantly enriched in the 308 transcripts with constitutive, high-level expression.

Supplemental Figure 6. Characterization of differentially expressed genes identified by DESeq during erythropoiesis. The heatmap shows a global view of the differential expression genes that passed FDR < 0.05 by DESeq. Expression levels (based on raw reads count) of the differentially expressed genes were normalized to those on day 4, and then binary logarithmic transformations of fold-change values were plotted to generate the heatmap by hierarchical clustering.

Supplemental Table 1: Total raw and uniquely mapped reads from two independent RNA-seq experiments

	Replicate 1		Replicate 2		Combination of two replicates	
	Raw	**Unique***	**Raw**	**Unique***	**Raw**	**Unique***
Day 4	68,650,800	52,071,902 (*76*)	87,060,828	66,017,834 (*76*)	155,711,628	118,089,736 *(76)*
Day 8	78,103,256	58,052,730 (*74*)	64,153,532	46,817,708 (*73*)	142,256,788	104,870,438 *(74)*
Day 11	73,378,172	53,799,010 (*73*)	68,406,088	46,551,948 (*68*)	141,784,260	100,350,958 *(71)*
Day 14	64,085,866	45,776,214 (*71*)	70,946,144	45,305,334 (*64*)	135,032,010	91,081,548 *(67)*

** Numbers in parentheses indicate the percentages of uniquely mapped reads*

Supplemental Table 2: Numbers and percentages of previously annotated genes and transcripts identified in expression profile analyses

	Gene		Transcript	
	Number	Percentage	Number	Percentage
Total in human	23,135	100	38,534	100
Day 4	17,501	75.6	26,833	69.6
Day 8	17,149	74.1	26,158	67.9
Day 11	16,832	72.8	25,452	66.1
Day 14	16,907	73.1	25,452	66.1
Union	18,770	81.1	30,140	78.2
Intersection	15,368	66.4	21,923	56.9

Supplemental Table 4: Top enriched Gene Ontology (GO) terms for Cluster 1, 4 and 7 transcripts

	GO description	*P*-value
Cluster 1	hemoglobin complex	1.03E-9
	nucleosome assembly	3.17E-9
	oxygen transport	2.36E-06
	cell death	2.68E-05
	heme metabolic process	2.92E-04
	Porphyrin and chlorophyll metabolism (KEGG pathway)	1.64E-04
	MAPK signaling pathway (KEGG pathway)	8.45E-04
Cluster 4	glycolysis	2.85E-05
	endocytosis	2.34E-04
	lipid metabolic process	2.68E-04
	Glycolysis / Gluconeogenesis (KEGG pathway)	1.40E-04
Cluster 7	GTPase regulator activity	7.86E-12
	membrane-bounded organelle	2.81E-09
	metal ion binding	7.41E-09
	immune system process	1.21E-07
	leukocyte activation	2.19E-06
	Fc gamma R-mediated phagocytosis (KEGG pathway)	3.44E-05

Supplemental Table 7: Top enriched Gene Ontology (GO) terms for Cluster 1 and 2 genes by DE_Seq

Clusters	Enriched GO terms	*P*-value
Cluster 1	cell cycle	5.21E-21
	erythrocyte differentiation	1.30E-07
	heme metabolic process	1.33E-06
	nucleosome assembly	1.92E-06
	Cell cycle (KEGG pathway)	3.39E-07
	Porphyrin and chlorophyll metabolism(KEGG pathway)	4.22E-06
Cluster 2	leukocyte activation	3.77E-10
	response to wounding	2.56E-09
	GTPase activator activity	5.01E-09
	defense response	5.87E-09
	lymphocyte activation	2.57E-08
	platelet alpha granule	5.59E-07
	Fc gamma R-mediated phagocytosis (KEGG pathway)	2.06E-07

Supplemental Table 8: Top enriched Gene Ontology (GO) terms for extremely differentially expressed transcripts

	GO description	*P*-value
Cluster 1	hemoglobin complex	8.80E-09
	oxygen transport	8.31E-07
	oxygen binding	3.35E-06
	iron ion binding	1.88E-05
	erythrocyte differentiation	3.99E-05
	response to chemical stimulus	1.56E-04
	nucleosome assembly	2.58E-04
	negative regulation of cell proliferation	3.52E-04
	heme binding	5.03E-04
Cluster 2	protein binding	1.26E-05
	cadmium ion binding	2.31E-04
	signal transduction	5.90E-04
	response to chemical stimulus	1.66E-03
	regulation of apoptosis	1.82E-03
Cluster 3	chemokine activity	5.76E-05
	chemotaxis	2.14E-03
	platelet alpha granule lumen	2.18E-03
	carbohydrate binding	2.52E-03
	Cytokine-cytokine receptor interaction (KEGG pathway)	2.45E-03
	Chemokine signaling pathway (KEGG pathway)	8.42E-03
Cluster 4	tube lumen formation	5.87E-04
	protein transport	2.77E-03
	positive regulation of protein kinase cascade	3.55E-03
Cluster 6	lymphocyte activation	5.09E-11
	immune response	8.84E-11
	T cell activation	1.55E-10
	leukocyte activation	2.05E-10
	positive regulation of immune system process	2.97E-09
	defense response	3.52E-09
	KEGG pathway	
	Leukocyte transendothelial migration	9.17E-07
	Cell adhesion molecules (CAMs)	4.50E-06
	Intestinal immune network for IgA production	7.61E-06
	Fc epsilon RI signaling pathway	1.02E-04
	Fc gamma R-mediated phagocytosis	1.46E-04
	Hematopoietic cell lineage	1.02E-03

Supplemental Table 9: qRT-PCR primers

Genes	Forward primer	Reverse primer
ACTG1	CCA AGG CCA ACA GAG AGA AG	CCT GGA TGG CCA CGT ACA TG
ACTN1	GCC AGC AAA GGC GTC AAA	TTC ACA TTC CCA TCC ACG ATT
AHSP	GGA TCT CAT TTC CGC AGG ATT G	CTG CTG CCT GTA ATA GTT GAT GT
HBA	GCA CGC TGG CGA GTA TGG	AAG TGC GGG AAG TAG GTC TTG GT
ATP2A3	TGG AGA ACC TGC AGT CCT TTA AC	GGT GCG TCG TTC ACT CCA T
BCL11A	TAT GCC CCG CAG GGT ATT TG	GCA GGT TAA AGG GGT TAT TGT CT
HBB	AAC TGT GTT CAC TAG CAA CCT CAA	GAG TGG ACA GAT CCC CAA AGG A
BLVRB	CAG GCT GTG ACT GAT GAC CAC	TCA CTG TGT ACG CCC CAG TTA
CA1	CCA AAG CTG CAG AAA GTA CTT GA	AAG GCA TCG TTG GAG TTC AG
CLC	GGC GAC CAC TTG CCT GTT T	CCT TCA TCT CAG TGT GGA AAT CC
DDN	TGG TGA TAG AAG TGA AGA CTA TTT CC	CCG CAC CTT GTC GAT GAA G
FAM178B	CCA GGA GCA ACA GCC AAA G	CAC AGG TAG CAG GCC TTG TG
HBG	GAT GCC ATA AAG CAC CTG GAT G	TTG CAG AAT AAA GCC TAT CCT TGA
GAPDH	CCA CAT CGC TCA GAC ACC AT	CCA GGC GCC CAA TAC G
GDF15	TTG CGG AAA CGC TAC GAG G	GCA CTT CTG GCG TGA GTA TCC
GYPC	GAC GAG AAG CCC CAA CAG	GTC TCC ATT CTG CCA TCC G
JUND	TCA TCA TCC AGT CCA ACG GG	TTC TGC TTG TGT AAA TCC TCC AG
LDHA	AGG TGG TTG AGA GTG CTT ATG AG	CCT TCG GAT TCT CCT TTT CTC TT
LDHB	TGG TAT GGC GTG TGC TAT CA	TCC ACA AGA GCA AGT TCA TCA G
LXRα	CCC TTC AGA ACC CAC AGA GAT C	GCT CGT TCC CCA GCA TTT T
MYB	CCC AAG TCT GGA AAG CGT CA	TTC GAT TCG GGA GAT AAT TGG C
NCL	AAC TAG AGA AAC CAA AAG GAA AAG ACA	CCT GAG TGA CTT TGT AAG GGA GAT TT
NME2	CCT CCG GGC CTC TGA AG	GGT CGG TCT TTC AGG TCA ATG
PCSK9	GCA CCT GCT TTG TGT CAC AGA	CTC GGC AGA CAG CAT CAT G
PLIN2	CCA TTT CTC AGC TCC ATT CTA CTG	TGA ATT TTC TGA TTG GCA CTA TAC A
PPIB	TGG ATA ATT TTG TGG CCT TAG C	CAC GAT GGA ATT TGC TGT TTT
PRDX2	GAC GCT TGT CTG AGG ATT ACG	AGG CCC CTG TAG GCA ATG C
RPS29	CCA CCC GCG AAA ATT CG	ACC GTG CCG GTT TGA AC
RPS5	ATG ACC GAG TGG GAG ACA G	CCT TCA CTG CAA TGT AAT CCT GC
RPS6	TGT CCG CCT GCT ACT GAG TAA	GCA ACC ACG AAC TGA TTT TCT C
RPS7	TGG AGA TGA ACT CGG ACC TCA	GTA CTA GCC GGA CTT GGA TTT TC
SLC7A5	GGA AGG GTG ATG TGT CCA ATC	CAC AAT GTT CCC CAC ATC CA
SRGN	CGG CTT GTC CTG GCT CTT G	CGT AGG ATA ACC TTG AAC TGA GGA T
TCP11L2	CTA AAT GCT GAC CCT CCT GAG T	GGT CGG ACA TAC CCT GCT G
TMSB10	GGC CAA GCT GAA GAA AAC G	CCT GCT CAA TGG TCT CTT TGG T
TR2	GCA GAC CAA CGG TGA TGT TTC	GGC TGT GCT CTC TCC AGG AT
TR4	AGA TGG GCA TGA AAA TGG AAT G	TGG TTT CTC CCG TTG CAC AT
18s	ACC GCA GCT AGG AAT AAT GGA	GCC TCA GTT CCG AAA ACC A

Supplemental Table 10: PCR primers for the nine novel intergenic transcripts analyzed

ID	Forward primer	Reverse primer
TCONS_00018850	AGA GGT GGC AGA TAT GCT GTT GGA	AGC AGC TGG CAT GAT GGG ATA GAT
TCONS_00040131	ATC ATG GTG GCA GAC AGA AGT GGA	TAA CTG GAA ATA CGG AGC CAG GCA
TCONS_00053016	TTC AAC TCT GGC CCT CAC AAT CCA	TTC CAG CAC TGG GAA ACT AGG GTT
TCONS_00004198	GAC TGC AGA ATG GAA ACA AGC CCA	AGA CTC CTG CCC ATA AAC ACC CAT
TCONS_00001777	ACC TAC AAG ATA TTC CTC TAC ACA	TCT TCA TCT GGA TCT ACA GGT GGG
TCONS_00057479	AGA TTC CAG TGC ACT TAT GAG G	CAG GAG TCA GAG TCA CTT TCA CAA GG
TCONS_00058467	GCT GTC ATG GTG CTT ATG AGG TGA	AGG TCT GAG ATT CCA TTC TGT CCC
TCONS_00014025	TGC CGT CCA GCA GGC CAC AA	TTC TGA GTG GAG CCT TGC ATC TGA
TCONS_00036982	GGC CAA GTT CGA CAC TGG TAA CCT	CAG CAA TGG TGA GGC AGA TAC CTT