CHAPTER 2

Materials and methods.

2.1. Culture and preparation of cell lines

The cell lines CHRF-288-11 and K562 were kindly provided by Katrin Voss (Department of Haematology, University of Cambridge; NHS Blood and Transplant, Cambridge, UK).

CHRF-288-11. The human megakaryoblastic cell line CHRF-288 was originally established from a biopsy of a metastatic solid tumour in a 17 month old infant with acute megakaryoblastic leukaemia (Witte et al., 1986). The cloned cell line (designated CHRF-288-11) exhibits markers characteristic of megakaryocytes and platelets. The cells also produce both basic fibroblast growth factor (bFGF) and transforming growth factor-β (TGF-β) (Fugman et al., 1990; Saito, 1997). CHRF-288-11 cells were maintained in RPMI-1640 medium [Sigma-Aldrich] supplemented with 20% horse serum (heat inactivated) [Invitrogen] and 1% L-glutamine-penicillin-streptomycin solution [Sigma-Aldrich].

K562. The human cell line K562 was established from a patient with chronic myeloid leukaemia (CML) in acute blast crisis (Lozzio & Lozzio, 1975). The glycoprotein pattern of K562 cells shows a striking similarity to that observed in normal erythrocytes (Andersson et al., 1979; Koeffler & Golde, 1980; Tabilio et al., 1983). K562 cells were maintained in RPMI-1640 medium [Sigma-Aldrich] supplemented with 10% foetal bovine serum (non-heat inactivated) [Biosera], 2 mM GlutaMAX-I [Invitrogen] and 1% L-glutamine-penicillin-streptomycin solution [Sigma-Aldrich].

Subculturing. Cells were taken from liquid nitrogen storage and thawed for 2 min at 37°C in a water bath. Cells were immediately removed from the vile with a sterile pipette and diluted in 20 ml of fresh growth medium. To remove dimethyl sulfoxide (DMSO), the cell suspension was centrifuged for 5 min at 72xg. Cells were resuspended in supplemented growth medium to ~$5x10^5$ cells/ml. CHRF-288-11 and K562 cells were grown at 37°C, 5% CO_2 and 100% humidified atmosphere. The cell suspension was diluted to ~$1x10^5$ cells/ml and fed or subcultured every 2–3 days. All cell culture work was performed in a class II microbiological safety cabinet.

Cell freezing. Cells were collected at early passage numbers and concentrated to $5x10^6$ cells/ml in fresh growth medium supplemented with 10% DMSO [Sigma-Aldrich]. The cell suspension was aliquoted into 1.2 ml cryopreservation vials [Nunc] (1 ml of cell suspension per tube) and placed in a freezing container [Mr. Frosty, Nalgene] at -80°C overnight. The freezing container was filled with isopropyl alcohol (IPA) [VWR BDH Prolabo], which ensures the 1°C/min-cooling rate required for successful cryopreservation of cells. Finally, vials were stored in liquid nitrogen until use.

2.2. Isolation, culture and preparation of primary cells

Ethics statement. Umbilical cord blood was obtained after informed consent under a protocol approved by the NHS Cambridgeshire Research Ethics Committee (REC 07/MRE05/44).

Monocyte isolation. Monocytes (MOs) were isolated from residual leukocytes obtained following apheresis platelet collections from NHS Blood and Transplant donors. Each sample (7.5 ml) was diluted 1:2 with PBE buffer (PBS [Sigma-Aldrich] at pH=7.2, 2 mM EDTA [Sigma-Aldrich] and 0.5% BSA [Sigma-Aldrich]) and gently layered onto the membrane of a 50 ml Leucosep tube [Greiner Bio-One]. By using these tubes, optimal separation of peripheral blood mononuclear cells (PBMCs) from human whole blood can be achieved by means of density gradient centrifugation using a porous, high-grade polyethylene barrier. Samples were centrifuged for 15 min at room temperature (RT) and 800xg. The PBMC layer was transferred into a fresh 50 ml tube. PBMCs from different Leucosep tubes were pooled, washed three times with 25 ml PBE buffer and centrifuged for 5 min at RT and 500xg. PBMCs were counted, diluted to $1x10^8$ cells/ml with PBE buffer and transferred into 5 ml polystyrene round-bottom tubes [BD Biosciences]. MO isolation was performed using the EasySep Human CD14 Positive Selection Kit [StemCell Technologies] according to the manufacturer's instructions. MOs, which strongly express the CD14 antigen, were targeted with tetrameric antibody complexes recognising CD14 and dextran-coated magnetic particles. Labelled cells were then separated using magnets, without the use of columns.

Megakaryocyte and erythroblast culture. Cord blood of newborns was collected into cord blood collection bags [MacoPharma]. $CD34^+$ haematopoietic progenitor cells (HPCs) were purified using the CD34 MicroBead Kit [Miltenyi Biotec] following the manufacturer's instructions. First, the $CD34^+$ cells were magnetically labelled with CD34 microbeads. Then, the cell suspension was loaded onto a column that is placed in a magnetic field, retaining the $CD34^+$ cells within the column. Purity (92–98%) and viability of HPCs were tested by flow cytometry. For *in vitro* differentiation of HPCs into megakaryocytes (MKs), 150,000 cells/ml/well were seeded in serum-free medium [CellGro SCGM, CellGenix] supplemented with 50 ng/ml human recombinant thrombopoietin (rhTPO) [CellGenix] and 10 ng/ml interleukin-1β (rhIL-1β) [Miltenyi Biotech]. To differentiate HPCs into erythroblasts (EBs), 5,000 cells/ml/well were seeded in serum-free medium supplemented with 6 U/ml erythropoietin (rhEPO) [R&D Systems], 10 ng/ml interleukin-3 (rhIL-3) [Miltenyi Biotech] and 100 ng/ml stem cell factor (rhSCF) [R&D Systems]. Cells were cultured for 7–10 days at 37°C and 5% CO_2. On the day of harvest, a cell aliquot was stained with 0.2% Trypan Blue [Sigma-Aldrich] and live cells counted using a haemocytometer [InCyto C-Chip, VWR International].

Cell morphology and flow cytometric analysis. For cell morphological analysis (**Figure 2-1**), aliquots of 50,000 cells were centrifuged onto a glass slide for 5 min at RT and 400xg and stained with modified Wright's stain using an automated slide stainer [HemaTek 1000, Miles Laboratories]. Stained cytospins were microscopically analysed [Axiovert 40 CFL, AxioCam HSc and AxioVision v4.5, Carl Zeiss MicroImaging]. Aliquots of 300,000 cells were used for flow cytometry. MOs were stained with human anti-CD14-PE clone TUK4 and anti-CD45-FITC clone c29/33 [Alere], as well as FITC and PE mouse monoclonal IgG1 isotype control [BD Biosciences]. After antibody incubation, 500 µl PBE buffer (PBS [Sigma-Aldrich] at pH=7.2, 2 mM EDTA [Sigma-Aldrich] and 0.5% BSA [Sigma-Aldrich]) and 5 µg/ml 7-amino actinomycin D (7-AAD) [Invitrogen] were added. Flow cytometric analysis of MKs and EBs was performed as previously described (Macaulay et al., 2007; Tijssen et al., 2011), using the following antibodies: human anti-CD41a-APC clone HIP8, anti-CD42a-FITC clone ALMA.16, anti-CD235a-FITC clone GA-R2 (HIR2) and anti-CD34-PE clone 581 [BD Biosciences]. All samples were analysed on the CyAn ADP 9-Color flow cytometer using the software Summit v4.3.02 [Beckman Coulter].

Ploidy stain of megakaryocytes. An aliquot of $1x10^6$ MKs was fixed with 70% (w/v) ethanol [Sigma-Aldrich] for 30 min at RT, washed once with PBE buffer (PBS [Sigma-Aldrich] at pH=7.2, 2 mM EDTA [Sigma-Aldrich] and 0.5% BSA [Sigma-Aldrich]) and stained with human anti-CD41a-APC clone HIP8 [BD Biosciences] or matched isotype control, as described above. After centrifugation, cells were resuspended in 500 µl staining buffer (465 µl PBE buffer, 5 µl 10% Tween-20 [Sigma-Aldrich], 5 µl of 10 mg/ml RNase A [Sigma-Aldrich] and 25 µl propidium iodide [Sigma-Aldrich]). After incubation for 30 min at 37°C, DNA content was determined using flow cytometry.

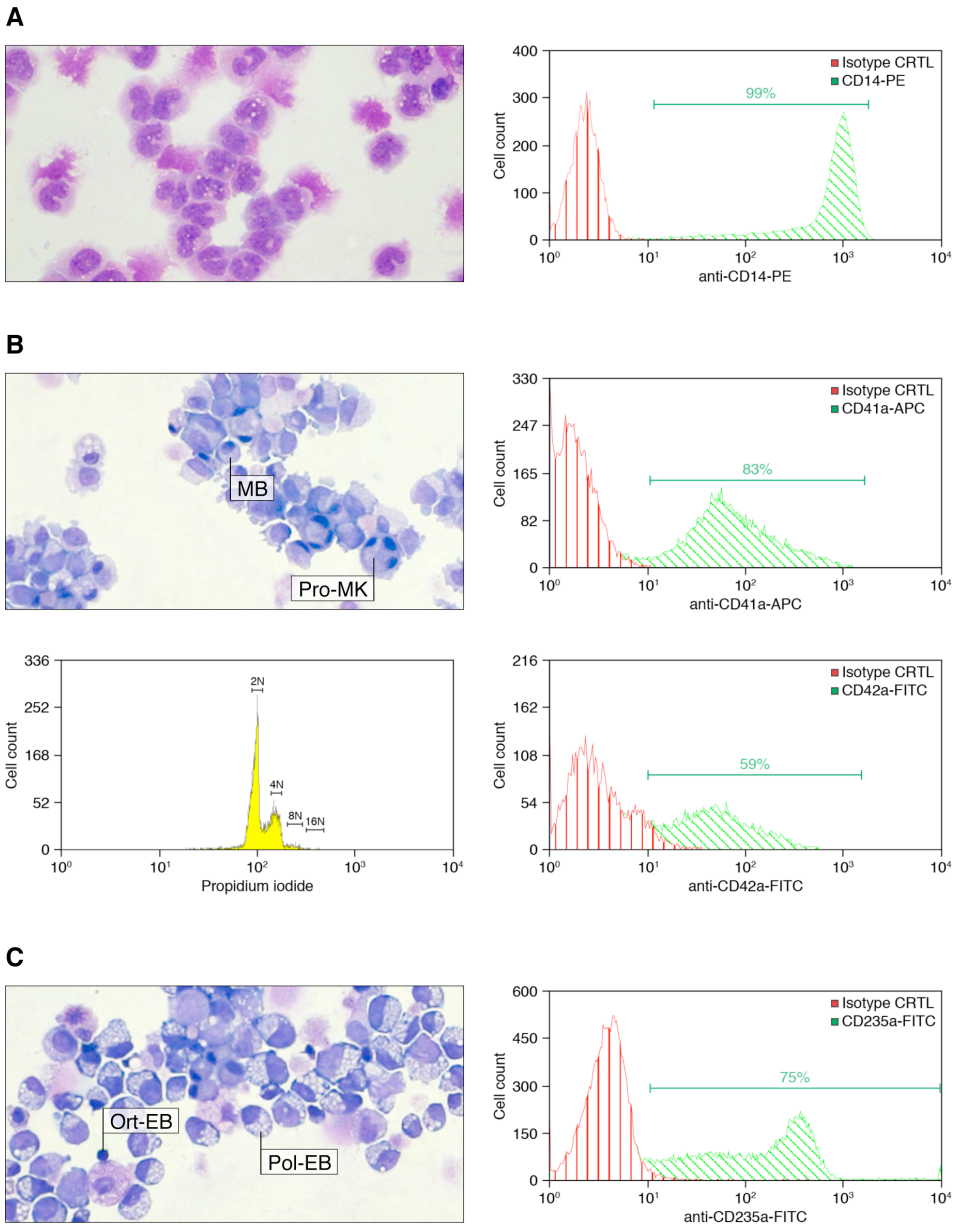

Figure 2-1. Characterisation of primary human MOs, MKs and EBs. Left panel: Representative images of stained cytospins (magnification, x40). Right panel: Gated cell populations with cell type-specific markers are shown in green, with corresponding IgG control populations in red. Indicated ranges represent average expression of gated cell populations of independent biological triplicates. For all cell cultures, more than 90% of cells tested negative for CD34 expression and corresponding isotype controls, as determined by flow cytometric analysis. Markers of other lineages were not detected. (**A**) Mature MOs were isolated based on morphological evaluation through microscopic analysis of stained cytospins (Goasguen et al., 2009). Isolated cells showed morphological characteristics of mature MOs, i.e. lobulated nucleus, condensed chromatin, occasional granules but no visible nucleolus. Histograms of flow cytometric analysis determined that 98% (range, 98–99%) of isolated cells expressed CD14 (and CD45, data not shown). (**B**) After megakaryocyte culture, analysis of stained cytospins and modified Wright's stain showed that the majority of cells were progenitor cells (megakaryoblasts, MB). MK cultures also contained pro-megakaryocytes (Pro-MK, horseshoe-shaped nucleus) and mature MKs (multinucleated) (Zeuner et al., 2007). Flow cytometric characterisation

revealed that 77% (range, 71–83%) of cells expressed CD41a and 44% (range, 28–59%) also expressed CD42a, which is only expressed by mature MKs. Ploidy analysis showed that 24.5% of MKs were 4N or higher. (C) In erythroblast cultures, cells were predominantly at a polychromatic (Pol-EB) and orthochromatic (Ort-EB) stage of differentiation, as determined by microscopic analysis (Panzenböck et al., 1998). Flow cytometric analysis showed that 71.6% (range, 68–75%) of cells expressed CD235a.

2.3. Formaldehyde-assisted isolation of regulatory elements (FAIRE)

Formaldehyde cross-linking. Cells (as described in **Table 2-1**) in fresh growth medium were transferred into a 150 mm x 25 mm cell culture dish or a 50 ml tube, and 37% formaldehyde [Merck Calbiochem] was directly added to the cell suspension to a final concentration of 1%. The cells were incubated at RT with gentle shaking on an orbital shaker [SO3, Stuart]. The cross-linking time across experiments varied and is reported in **Table 2-1**. To quench the fixation, 2.5 M glycine [AppliChem] was added to a final concentration of 125 mM, and the cell suspension was shaken for 5 min. The cells were collected by centrifugation for 5 min at 340xg. Cell pellets were washed with cold 1x PBS to remove all residual media. Cells were collected for 5 min at 340xg.

Cell lysis for optimisation and FAIRE-chip experiments. The cell pellet was resuspended in lysis buffer L1 (50 mM HEPES-KOH [Sigma-Aldrich] at pH=7.5, 140 mM NaCl [VWR BDH Prolabo], 1 mM EDTA [Amresco] at pH=8.0, 0.50% Igepal CA-630 [USB Corporation], 0.25% Triton X-100 [Sigma-Aldrich] and 10% glycerol [VWR BDH Prolabo]) to a concentration of $10x10^6$ cells/ml, and incubated for 10 min on ice. The cell suspension was aliquoted into 1.5 ml tubes (500 µl of lysate per tube), and cells were collected for 5 min at 4°C and 1,300xg. Next, the pellet was resuspended in 500 µl of lysis buffer L2 (200 mM NaCl [VWR BDH Prolabo], 1 mM EDTA [Amresco] at pH=8.0, 0.5 mM EGTA [Merck Calbiochem] at pH=8.0 and 10 mM Tris-HCl [Sigma-Aldrich] at pH=8.0), and incubated for 10 min at RT. The cell suspension was spun down for 5 min at 4°C and 1,300xg. The pellet was then resuspended in 300 µl of lysis buffer L3 (100 mM NaCl [VWR BDH Prolabo], 1 mM EDTA [Amresco] at pH=8.0, 0.5 mM EGTA [Merck Calbiochem] at pH=8.0, 10 mM Tris-HCl [Sigma-Aldrich] at pH=8.0, 0.1% Na-deoxycholate [Sigma-Aldrich] and 0.5% (w/v) N-lauroylsarcosine sodium salt [Sigma-Aldrich]). Before use, 2x EDTA-free Protease Inhibitor [Complete Mini, Roche] were added to each 25 ml of L3, and immediately sonicated.

Cell lysis for FAIRE-seq experiments. The cell pellet was resuspended in 5 ml of ice-cold PBS supplemented with 1x EDTA-free Protease Inhibitor [Complete Mini, Roche]. The sample was spun for 6 min at 4°C and 249xg. The cell pellet was then resuspended again in 2 ml of lysis buffer (10 mM Tris

[Thermo Fisher Scientific] at pH=8.0, 10 mM NaCl [VWR BDH Prolabo], 1x EDTA-free Protease Inhibitor [Complete Mini, Roche] and 0.2% Tergitol solution [Type NP-40, Sigma-Aldrich]. The sample was incubated for 10 min on ice. To recover the nuclei, the sample was spun for 5 min at 4°C and 1,083xg. The supernatant was removed and the nuclei resuspended in 2 ml of nuclei lysis buffer (50 mM Tris [Thermo Fisher Scientific] at pH=8.1, 10 mM EDTA [Thermo Fisher Scientific], 1x EDTA-free Protease Inhibitor [Complete Mini, Roche] and 1% SDS [VWR BDH Prolabo]). The sample was then incubated for 10 min on ice. Finally, 2 ml of dilution buffer (20 mM Tris [Thermo Fisher Scientific] at pH=8.1, 2 mM EDTA [Thermo Fisher Scientific], 150 mM NaCl [VWR BDH Prolabo], 1x EDTA-free Protease Inhibitor [Complete Mini, Roche], 1% Triton X-100 [Sigma-Aldrich] and 0.01% SDS [VWR BDH Prolabo]) was added, and immediately subjected to sonication.

Sonication. The samples were aliquoted to a final volume of 300 µl (~2–5x10^6 cells) per 1.5 ml tube, i.e. four and six aliquots for FAIRE-chip and FAIRE-seq experiments, respectively. Chromatin was subjected to sonication cycles of 30 sec at high pulse (200 W) followed by 30 sec of rest using the Bioruptor UCD-200 [Diagenode]. The sonication time varied across experiments and is reported in **Table 2-1**. A temperature of ~4°C was maintained. For optimisation experiments (**Table 2-1**; discussed in **Section 3.2**), aliquots of 2 µl were taken after respective sonication cycles to monitor sonication efficiency. Prior to analysis with a 2100 Bioanalyzer [Agilent Technologies], DNA-protein cross-links were reversed (see next paragraph). Finally, the lysate was cleared of cellular debris by spinning for 5 min at 4°C and 15,000xg, and the supernatant transferred into a new tube.

Reverse cross-linking. For optimisation experiments, samples were incubated for 6 hr at 65°C in a thermocycler [Thermomixer 5436, Eppendorf] prior to analysis using a 2100 Bioanalyzer [Agilent Technologies].

Analysis of DNA fragment length. For optimisation experiments, 1 µl of sample from each time point was applied to a 2100 Bioanalyzer [Agilent Technologies], using a DNA 1000 Chip or DNA 7500 Chip [Agilent Technologies] according to the manufacturer's protocol. Data was analysed using the software 2100 Expert [Agilent Technologies]. Experiments were performed with biological and technical replicates.

Phenol-chloroform extraction. An equal volume (300 µl) of phenol-chloroform-isoamyl alcohol (25:24:1) [Sigma-Aldrich] saturated with 10 mM Tris at pH=8.0 and 1 mM EDTA was added to the lysate. The mixture was vortexed, centrifuged for 5 min at 4°C and 12,000xg, and the aqueous phase transferred to a new tube. Then, 300 µl of TE buffer (10 mM Tris-HCl [Sigma-Aldrich] at pH=7.5 and

1 mM EDTA [Sigma-Aldrich]) was added to the organic phase, vortexed and centrifuged for 5 min at 4°C and 12,000xg. The aqueous phase was extracted and combined with the first extraction. To remove residual protein, an additional round of extraction was performed by adding 300 µl of phenol-chloroform-isoamyl alcohol to the combined aqueous fraction, followed by thorough mixing, centrifugation and retention of the aqueous phase. Then, 400 µl of chloroform-isoamyl alcohol (24:1) [Sigma-Aldrich] was added. The tube was vortexed, the two phases separated by centrifugation for 5 min at 4°C and 12,000xg, and the aqueous phase (400 µl) retained.

DNA precipitation. One-tenth volume (40 µl) of 3 M sodium acetate [VWR BDH Prolabo] at pH=5.2 and 1 µl of 20 mg/ml glycogen [Roche] were added to the mixture, and the tube was mixed by inverting. Two volumes (800 µl) of 95% ethanol [VWR BDH Prolabo] were added to the mix by inverting, and the reaction was incubated at 4°C overnight. Precipitated DNA was collected for 30 min at 4°C and 15,000xg, and the pellet washed with 500 µl of 70% ice-cold ethanol [VWR BDH Prolabo] by centrifugation for 10 min at 4°C and 15,000xg. The pellet was dried for ~10 min at RT in a SpeedVac [Concentrator 5301, Eppendorf], and resuspended in 25 µl of 10 mM Tris-HCl [Sigma-Aldrich] at pH=7.5. The aliquots were combined to form aliquots of 50 µl each. Next, 1 µl of 200 µg/ml RNase A [ICN Biomedicals] was added to the mix and incubated for 1 hr at 37°C. Finally, DNA was purified using the MinElute PCR Kit [Qiagen] according to the manufacturer's protocol. DNA was eluted in 2x10 µl of EB Buffer.

Table 2-1. Overview of experimental parameters applied in FAIRE experiments.

Parameter	Optimisation			FAIRE-chip		FAIRE-seq
	FAIRE	*FAIRE*	*Reference*	*FAIRE*	*Reference*	
Cell number	$20x10^6$	$20x10^6$	$5x10^6$	$20x10^6$	$20x10^6$	$15x10^6$
Fixation	5 min	5, 8, 12 min	n/a	8, 12 min	n/a	12 min
Sonication	9, 12, 14 min	12 min	4, 6, 8, 9 min	12 min	9 min	12 min
Cell type(s)	CHRF-288-11			CHRF-288-11, K562		CHRF-288-11, MKs, EBs, MOs

2.4. Detection and analysis using DNA tiling microarrays

Sample labelling. Precipitated DNA recovered from cross-linked cells ('FAIRE sample') and uncross-linked cells ('reference sample') was labelled with Cy5 and Cy3 dye, respectively. Sample labelling was performed using the Dual-Color DNA Labeling Kit [all components, Roche NimbleGen] according to

the manufacturer's protocol (Roche NimbleGen Arrays User's Guide, ChIP-chip Analysis v4.1). Both Cy3- and Cy5-Random Nonamer Primers were diluted in 462 µl of Random Primer Buffer supplemented with β-Mercaptoethanol [Sigma-Aldrich]. The FAIRE and reference samples were placed in separate 200 µl thin-walled PCR tubes [Ambion]. For FAIRE and reference samples, each 80 µl reaction contained 40 µl of purified DNA (**Table 2-2 A**) and 40 µl of diluted Cy5- and Cy3-Random Nonamers, respectively. The samples were heat denatured for 10 min at 98°C in a thermocycler [Thermomixer 5436, Eppendorf], immediately quick-chilled in an ice-water bath and incubated for 2 min. Next, 20 µl of the dNTP/Klenow Master Mix was added on ice to each of the denatured samples to a final volume of 100 µl. The solution was mixed well by pipetting up and down 10 times. Subsequently, the mix was collected by centrifugation and incubated for 2 hr at 37°C in a thermocycler with heated lid, protected from light. The reaction was stopped by addition of 10 µl of 0.5 M EDTA. Then, 11.5 µl of 5 M NaCl was added to each sample. The mix was briefly vortexed, spun down and the entire contents transferred to a 1.5 ml tube containing 110 µl of 100% isopropanol. After vortexing, the mix was incubated for 10 min at RT, protected from light. The precipitate was collected by centrifugation for 10 min at 12,000xg. The supernatant was removed and the pellet rinsed with 500 µl of 80% ice-cold ethanol by centrifugation for 2 min at 12,000xg. The supernatant was removed and the pellet dried for ~10 min at 30°C in a SpeedVac [Concentrator 5301, Eppendorf], protected from light. The dried pellet was rehydrated in 25 µl of nuclease-free water. Finally, the sample was vortexed several times until the pellet was completely rehydrated. After labelling, the FAIRE and reference samples were quantitated using a NanoDrop spectrophotometer [ND-1000, Labtech]. Samples were analysed in the Nucleic Acid module, DNA-50 mode using the software ND-1000 v3.5.2 (**Table 2-2 B**).

Table 2-2. DNA quantity of FAIRE and reference samples before and after labelling with cyanine dyes.

Cell line	(A) Before labelling			(B) After labelling		
	Reference	*8 min*	*12 min*	*Reference*	*8 min*	*12 min*
CHRF-288-11	11.25 µg	0.79 µg	0.38 µg	21.72 µg	8.31 µg	6.20 µg
K562	11.75 µg	0.84 µg	0.59 µg	33.11 µg	11.62 µg	15.55 µg

For the array hybridisation reaction, 6 µg of both FAIRE and reference DNA were required (**Table 2-2 B**). Based on the determined concentration, the respective volumes of the FAIRE and reference samples were calculated and combined in a 1.5 ml tube. The content was dried for ~15 min at 30°C in a SpeedVac [Concentrator 5301, Eppendorf], protected from light.

Array hybridisation. The MAUI Hybridization System [BioMicro Systems] was set to 42°C and the temperature was allowed to stabilise for 3 hr. Hybridisation on 385K arrays was performed using the Hybridization Kit, Mixer X1 and Precision Mixer Alignment Tool (PMAT) [Roche NimbleGen] according to the manufacturer's protocol. The dried sample pellet was resuspended in 5 µl of DNase- and RNase-free water [Gibco]. The mix was vortexed well and spun down. Next, 13 µl of the Hybridization Solution Master Mix [Roche NimbleGen] was added to 5 µl of resuspended sample. The solution was vortexed, spun down and incubated for 5 min at 95°C, protected from light. The sample was incubated for 5 min at 42°C in the hybridisation system. The Mixer X1 was assembled with the 385K-feature slide using the PMAT. The assembly was placed in the slide bay of the hybridisation system, and 16 µl of the hybridisation mix was loaded into the fill port of the slide/mixer assembly. Finally, the sample was hybridised to the array for 20 hr at 42°C in mixing mode B.

Array washing. Arrays were washed using the Wash Buffer Kit [Roche NimbleGen] according to the manufacturer's protocol. To process the protocol without interruption and ensure high quality data, only one slide was washed at a time. After hybridisation, the mixer-slide assembly was removed from the hybridisation system and loaded in the Mixer Disassembly Tool that was immersed in 250 ml of warm Wash I (42°C). With the mixer-slide assembly submerged, the mixer was carefully peeled off the slide. The mixer was discarded and the slide quickly removed from the Mixer Disassembly Tool. The slide was gently agitated for 15 sec to quickly remove the hybridisation buffer. Subsequently, the slide was transferred into a slide container containing Wash I and agitated vigorously for 15 sec. During all wash steps, the microarray area of the slide was submerged at all times and not allowed to dry between wash steps. The slide was washed for an additional 2 min in Wash I with vigorous constant agitation. The slide was transferred to Wash II and washed for 1 min and subsequently to Wash III for 15 sec with vigorous, constant agitation. The slide was removed from Wash III and immediately dried for 2 min in a microarray centrifuge [Spectrafuge Mini, Labnet]. Residual moisture was removed by blow-drying the edges. Immediately after washing, the slide was scanned.

Array scanning. Arrays were scanned using a DNA Microarray Scanner [Agilent Technologies] and the software Scan Control v8.3.1 [Agilent Technologies] according to the manufacturer's protocol. Arrays were scanned with wavelengths of 532 nm and 635 nm for Cy3 and Cy5, respectively, PMT power of 100% and a pixel size of 5 µm.

Data processing. Experimental data were analysed using the software NimbleScan v2.5 [Roche NimbleGen]. The two-channel raw signal intensities were scaled between channels by subtracting the Tukey bi-weight mean for the $\log_2$-ratio values for all features from each $\log_2$-ratio value. This scaling

procedure accounts for differences in the signal intensities of the dyes by centring the data on zero. Since the experimental setup provided a two-colour array with the reference sample on the array, normalisation of data was not performed.

Peak calling. In order to find peaks in the scaled $\log_2$-ratio data, a sliding window was moved across each chromosome probe by probe. Within this window, each probe was tested if its $\log_2$-ratio was above a certain cut-off value. A peak was registered when the number of qualifying probes was above a set probe number within the sliding window. The genetic position of the identified peak was set from the start position of the first qualifying probe to the end position of the last qualifying probe. For each chromosome, the $\log_2$-ratio cut-off value was calculated as the percentage of a hypothetical maximum (P_{max}=arithmetic mean + 6x standard deviation). The peak finding process was repeated using a series of $\log_2$-ratio cut-off values from P_{start} to P_{end}. By using a hypothetical maximum rather than the overall maximum of the $\log_2$-ratios, the effects of outliers can be minimised (Lucas et al., 2007). The following settings for the peak finding analysis were applied: sliding window: 300 bp; min. probes>cut-off in peak=4; all probes in peak>cut-off=2; P_{start}=90%, P_{end}=15%, P_{step}=0.5, number of steps: 100. In all array experiments, the signal $\log_2$-ratio between the FAIRE and the reference sample showed a normal distribution, with an enrichment of FAIRE signal at the right end of the distribution.

Data visualisation. Log_2-ratio and peak data sets were displayed as UCSC Genome Browser (http://genome.ucsc.edu/) custom tracks.

Data availability. The FAIRE microarray data sets are available online in the Gene Expression Omnibus (GEO) database under accession number GSE25716.

2.5. Detection and analysis using high-throughput next-generation sequencing technology

Library preparation and sequencing. FAIRE DNA was processed following the Illumina paired-end library generation protocol. Genomic libraries derived from MO extractions and CHRF-288-11 cells were sequenced on Illumina HiSeq 2000 with 50 bp and 75 bp paired-end reads, respectively. Libraries derived from EB and MK cultures were sequenced on Illumina GAIIx with 54 bp paired-end reads. All FAIRE sample libraries were prepared and sequenced at the Wellcome Trust Sanger Institute by the library-making and sequencing core groups, respectively.

Sequence data processing. Raw sequence reads were aligned to the human reference sequence (NCBI build 37) using the algorithm Stampy (Lunter & Goodson, 2011). Reads were realigned around known insertions and deletions (The 1000 Genomes Project Consortium, 2010), followed by base quality recalibration using the Genome Analysis Toolkit (GATK) (McKenna et al., 2010). Duplicates were flagged using the software Picard (http://picard.sourceforge.net/) and excluded from subsequent analyses. The raw sequence files from two independent FAIRE experiments in K562 cells were obtained from the ENCODE Project (GEO accession number GSM864361), and remapped as described above. An overview of the sequencing statistics is provided in **Table 2-3**.

Table 2-3. Overview of sequencing statistics. Summary of the total number of DNA sequence reads mapped to the human reference sequence (NCBI build 37), listed for each sample. The following exclusion criteria were applied to sequence reads: mapped with quality below 30; duplicated; and mapped to mitochondrial DNA and unplaced chromosomes. For paired-end data sets, reads not properly paired and paired farther than 1 kb were also excluded.

Cell type	Samples	Number of mapped reads		Percentage of total reads
		Before filtering	*After filtering*	
EB	Indiv. A	40,823,840	29,202,677	71.5%
	Indiv. B	48,350,482	40,542,641	83.9%
MK	Indiv. B	43,292,704	36,376,461	84.0%
	Indiv. C	48,365,225	35,697,504	73.8%
MO	Indiv. D	250,497,524	201,993,051	80.6%
	Indiv. E	225,495,458	184,262,652	81.7%
K562	Repl. 1	59,913,440	43,401,021	72.4%
	Repl. 2	59,741,112	40,694,874	68.1%
CHRF	Repl. 1	58,867,716	52,438,908	89.1%
	Repl. 2	60,088,050	53,769,761	89.5%

Peak calling and normalisation. Regions of enrichment (peaks) were determined using the software F-Seq v1.84 (Boyle, Guinney, et al., 2008). A feature length of L=600 bp and two different standard deviation thresholds of T=6.0 ('moderate') and T=8.0 ('stringent') over the mean across a local background were applied. In order to reduce false positive peak calls, regions of collapsed repeats were removed, as described in Pickrell et al., 2011, applying a threshold of 0.1% (http://eqtl.uchicago.edu/Masking). Four equally spaced bins (in $\log_{10}$-transformed peak score units) were defined between the 1st and 99th percentile of the peak score distribution. For comparison of cell type-specific chromatin profiles, all read fragments were merged into one data set for each cell type. Then, peaks were called as described. For the K562 single-end sequencing data set, the mode of the peak width distribution was

adjusted to the mean of the modes across all non-K562 cell types. **Tables 2-4** and **2-5** give an overview of the final peak data sets. ChIP-seq data sets in primary MKs were obtained from Tijssen et al., 2011 (GEO accession number GSE24674). The peak coordinates were remapped to hg19 (minimum ratio of bases that must remap: 0.95) using the Lift-Over tool v1.0.3 of the web-based platform Galaxy (http://main.g2.bx.psu.edu/).

Table 2-4. Overview of FAIRE peak statistics.

Cell type	Samples	(A) Number of FAIRE peaks		(B) Number of merged FAIRE peaks	
		Moderate (T=6.0)	*Stringent (T=8.0)*	*Moderate (T=6.0)*	*Stringent (T=8.0)*
EB	Indiv. A	67,395	22,754	96,741	37,252
	Indiv. B	82,761	31,553		
MK	Indiv. B	86,270	28,543	148,124	49,364
	Indiv. C	81,622	26,845		
MO	Indiv. D	55,512	16,399	101,034	34,135
	Indiv. E	43,664	17,669		
K562	Repl. 1	84,356	41,638	122,463	67,832
	Repl. 2	84,991	43,981		
CHRF	Repl. 1	128,184	60,970	222,424	109,610
	Repl. 2	123,053	63,538		

Table 2-5. Overview of the number of FAIRE peaks for each intensity bin. Shown are the merged peak data sets, called with the stringent F-Seq threshold (T=8.0). Peaks were filtered based on the criteria described above, and did not count to the total number of peaks.

	EB	MK	MO	K562	CHRF
Bin 1	28,415	42,206	21,465	33,008	80,482
Bin 2	5,235	4,407	7,745	20,612	17,039
Bin 3	2,563	1,860	3,565	9,974	8,357
Bin 4	1,039	891	1,360	4,238	3,732
Total number	37,252	49,364	34,135	67,832	109,610
Total filtered	*762*	*1,008*	*698*	*1,694*	*2,238*

Data visualisation. The coverage profile on the combined data was created using the R packages ShortRead (Morgan et al., 2009) and rtracklayer (Lawrence, Gentleman, et al., 2009). Coverage and peak data sets were displayed as UCSC Genome Browser custom tracks.

Data availability. All FAIRE sequencing data sets are available online in the GEO database under accession number GSE37916.

2.6. Annotation of NDRs and statistical analyses

Hierarchical cluster analysis and bootstrapping. First, a union set of all peaks across all samples was created. Next, a vector of binary values for each sample *s* was defined, whereby the length of the vector is given by the total number of peaks in the union set and is therefore the same for all samples. Position *i* in the vector for sample *s* was set to a value of one, if the peak *i* in the union peak set overlaps with a peak in sample *s*. If there was no overlap with a peak in sample *s*, position *i* was set to zero. From these vectors, bin-specific vectors were constructed based on the binning scheme described in **Section 2.5**. For each bin and sample, a vector was defined where all entries with a peak score not between the lower and upper peak scores defined for that bin, were set to zero. Then, the R package Pvclust (Suzuki & Shimodaira, 2006) was used to perform a bootstrapped hierarchical cluster analysis of the samples based on these binary vectors, using the 'binary' distance measure and the 'complete' method for defining the clusters (Suzuki & Shimodaira, 2006). Here, 1,000 bootstrap samples were applied. All analyses were carried out in the R/Bioconductor environment.

Annotation of NDRs using GREAT. The ontology of genes flanking FAIRE peaks was analysed using the Genomic Regions Enrichment of Annotations Tool (GREAT) v1.8.2 (McLean et al., 2010) with the following parameters: association rule: single nearest gene; 1 Mb maximal extension; curated regulatory domains included. The genomic distances between FAIRE peaks and transcription start sites were exported from GREAT.

Enrichment analysis using bootstrapped quantile distributions. The association analysis for the eight quantitative traits was performed by imputation from the Phase II HapMap panel (The International HapMap Consortium, 2007). To improve the coverage, for each Phase II HapMap SNP it was determined which 1000 Genomes SNPs (interim phase I release of June 2011) within a distance of 50 kb had an r^2 of at least 0.95 with the imputed Phase II HapMap SNPs. For each trait, the *P*-value of the Phase II HapMap SNP from the meta-analysis to the 1000 Genomes SNP was assigned. To prevent chance inflation from LD and to obtain confidence estimates, 500 bootstrap samples were created from this set of 1000 Genomes SNPs by randomly removing SNPs until the genomic distance between remaining SNPs was at least 50 kb. For each trait, the mean genomic inflation at the 0.005 quantile was inferred from these bootstrap samples. This genomic inflation factor provides a baseline genomic

inflation factor. It should be noted that corrections for population stratification and covariates have already been performed in the original meta-analyses (Gieger et al., 2011; van der Harst et al., 2012). Using the same pruning approach, for each of the three peak sets ('merged', 'intersected' and 'cell type-specific'), 500 bootstrap samples were generated using only those 1000 Genomes SNPs that are located in a peak from the respective peak set. Finally, the relative genomic inflation factors (as reported in **Figure 4-7**) were calculated as the ratio between the baseline genomic inflation factor and the genomic inflation factor calculated for SNPs located in peaks from a given peak set. The 0.005 quantile provides a trade-off between highlighting differences in enrichment across different cell types and reducing uncertainty in the estimates of the relative genomic inflation factors. All analyses were carried out in the R/Bioconductor environment.

Canonical pathway analysis. Genes were subjected to the core analysis module of the software Ingenuity IPA v12402621 (http://www.ingenuity.com), and analysed using the following parameters: reference set: Ingenuity Knowledge Base (genes only); relationship to include: direct and indirect; filter: only molecules and/or relationships where species=human and confidence=experimentally observed. Benjamini-Hochberg multiple test-corrected *P*-values are reported.

Overlap of NDRs with association loci and significance analysis. For each association locus, candidate functional SNPs were selected by identifying all biallelic SNPs with an $r^2>0.8$ and within 1 Mb of the sentinel SNP in the European samples of the 1000 Genomes Project data set (interim phase I release of June 2011). To establish whether the association of a locus could potentially be of regulatory origin, it was determined if at least one candidate functional SNP overlapped with a FAIRE peak. Since this analysis is sensitive to the number of peaks, the overlap was carried out for successively increasing number of peaks by considering peaks with decreasing peak height. As more peaks are considered, the chance of finding an overlap increases. Therefore, the significance of the findings was estimated by resampling. Of both 68 and 75 loci associated with platelet and erythrocyte phenotypes, respectively, 100,000 sets were drawn from the same SNPs onto which the GWA data was imputed ($\sim 2.5 \times 10^6$ SNPs from Phase II HapMap), while preserving the distribution of allele frequencies and the number of loci that overlapped with FAIRE peaks. This was repeated (100,000 permutations) for each successive increase in the number of FAIRE peaks. All analyses were carried out in the R/Bioconductor environment.

2.7. Gene expression analysis during *in vitro* differentiation of cord blood-derived HPCs

Experiments and statistical analyses were performed as described in Gieger et al., 2011. Briefly, MKs and EBs were differentiated from cord blood-derived HPCs as described in **Section 2.2**. Time points were taken at days 3, 5, 7, 9, 10 and 12. Whole-genome gene expression levels were measured using Illumina HumanWG-6 v3 Expression BeadChips. Expressed probes were selected based on stringent thresholds and the slope of expression was determined using standard linear regression. The single closest ENSEMBL transcript (release 64) with an HGNC symbol was assigned to every FAIRE peak.

2.8. H3K4me1 and H3K4me3 ChIP-seq

MKs and EBs were differentiated from cord blood-derived HPCs as described in **Section 2.2**. ChIP assays were performed as described in Forsberg et al., 2000, using rabbit polyclonal antibodies against H3K4me1 [ab8895, Abcam] and H3K4me3 [07-473, Millipore]. Chromatin-immunoprecipitated DNA was sequenced on Illumina GAII with 54 bp single-end reads. Sequence reads were aligned using the algorithm BWA (Li & Durbin, 2009). Areas of enrichment were determined using the slice function of the R package IRanges (http://bioconductor.org/packages/2.10/bioc/html/IRanges.html).

2.9. Sanger sequencing of selected NDRs

Ethics statement. All human subjects were recruited with appropriate informed consent in Cambridgeshire and enrolled in the Cambridge BioResource (http://www.cambridge-bioresource.org.uk/).

Capillary sequencing. DNA samples from a total of 643 individuals of Northern European ancestry were subjected to capillary DNA sequencing of the targeted locus at chromosome 7q22.3. Sequencing primer pairs for two sequence-tagged sites were designed using the web-based tool Primer3 (Rozen & Skaletsky, 2000), and are reported in **Table 2-6**.

Table 2-6. Sanger sequencing primer pairs for two sequence-tagged sites at chr7q22.3. Genomic coordinates of the PCR products were based on the human reference genome, build hg18.

Forward primer sequence	Reverse primer sequence	Genomic position	Amplicon
TGGAAAATTACAAAAGTCCCAAA	GAGAAAGGATCATGAGGGAGAA	chr7:106,159,328–106,159,998	671 bp
ACAAAAGTCCCAAAATTTCACA	GAGAAAGGATCATGAGGGAGA	chr7:106,159,337–106,159,998	662 bp

PCR products were applied to bi-directional sequencing using Big Dye chemistry on 3730 DNA sequencers [Applied Biosystems]. DNA amplification by PCR and capillary sequencing were performed at the Wellcome Trust Sanger Institute by members of the ExoSeq/ExoCan facility. Details of the sequencing protocol are described online (http://www.sanger.ac.uk/resources/downloads/human/exoseq.html).

Analysis. Pre-processed sequence traces were analysed using the semi-automated analysis software ExoTrace (http://www.sanger.ac.uk/resources/downloads/human/exoseq.html), developed at the Wellcome Trust Sanger Institute. Results of the SNP calling were displayed in a specific implementation of GAP4, which is part of the Staden Sequence Analysis Package (http://staden.sourceforge.net/). Potential SNP positions were indicated by the software and then reviewed manually.

2.10. Transcription factor binding site prediction

TRAP. Transcription factor binding sites were predicted using the transcription factor affinity prediction (TRAP) method (Thomas-Chollier et al., 2011) (sTRAP tool) with the following parameters: matrix: TRANSFAC v2010.1 (vertebrates); background: human promoters; multiple test correction: Benjamini-Hochberg.

MathInspector. In **Section 5.2**, transcription factor binding sites were predicted using the software MatInspector v8.01 (Cartharius et al., 2005). The following parameters were applied: matrix group: vertebrates; core=1.00; matrix=optimised+0.02; tissue: haematopoietic system. In **Section 6.2**, an updated MathInspector library version was used (v8.3). The same parameters were applied, except for the matrix group: general core promoter elements and vertebrates. In addition, a restriction on tissue type was omitted.

2.11. Electrophoretic mobility shift assay (EMSA)

Extraction of nuclear protein. Non-denatured, active nuclear proteins were purified from $5x10^6$ CHRF-288-11 cells with the NE-PER Nuclear and Cytoplasmic Extraction Reagents [Thermo Fisher Scientific] according to the manufacturer's protocol. The cell pellets were removed from -80°C storage and incubated for 5 min on ice. Cytoplasmic Extraction Reagent I (CER I) at a volume of 200 µl was added to the ~20 µl of packed cell volume. The sample was vigorously vortexed for 15 sec on the highest setting to fully suspend the cell pellet, and subsequently incubated for 10 min on ice. Next, 11 µl of Cytoplasmic Extraction Reagent II (CER II) was added, and the tube vortexed for 5 sec on the highest setting. The sample was incubated for 1 min on ice. The sample was then vortexed for 5 sec on the highest setting, and centrifuged for 5 min at 4°C and 16,000xg. After cell membrane disruption and release of cytoplasmic contents, the supernatant (cytoplasmic extract) was immediately transferred to a clean pre-chilled tube. The pellet, which contains intact nuclei, was resuspended in 100 µl of ice-cold Nuclear Extraction Reagent (NER). The sample was vortexed for 15 sec on the highest setting. The sample was placed on ice whilst continuing to vortex for 15 sec every 10 min, for a total of 40 min. Then, the sample was centrifuged for 10 min at 4°C and 16,000xg. The supernatant (nuclear extract) was immediately transferred to a clean pre-chilled tube. Extracts obtained with this protocol generally have less than 10% contamination between nuclear and cytoplasmic fractions. For every EMSA, fresh nuclear protein was prepared.

Probe design. Oligonucleotides were designed based on the genomic sequence surrounding each candidate functional SNP (**Table 2-7**). Oligonucleotides were prepared with a biotin tag at the 5'-end and without modification ('competitor'), for both alleles of the candidate SNP. In addition, unlabelled complementary strands were prepared for both alleles. All oligonucleotides (desalting purification) were provided by Sigma-Aldrich.

Table 2-7. EMSA probes.

Candidate SNP	Sequence of probe 5'→3'
rs342293	AGCCCTGTGGTTTTAATTAT[C/G]TTGAGGTTCAGGCTCA
chr1:145,507,646 (5'-UTR *RBM8A*)	AGTGTCTGAGCGGCACAGAC[G/A]AGATCTCGATCGAAGG
chr1:145,507,765 (intron *RBM8A*)	AGACGGCTGGTGGGAAGC[G/C]GGGAAGGTGCGAGAGAAGG
rs1006409	TTCCTTCTTTTCCTTTT[A/G]TGGTATGCATAGATATCA
rs1107479	CTGCCAAGGACGTCA[C/T]AGGCAGATGGAAGGAAGCTT
rs11731274	TGGCACACGCTGGTGGC[T/G]TTCCCCGGGCTCTCTGCT
rs11734099	GAGCTCCCTCCCTGGCCT[G/A]CCTGGCACACGCTGGTGGCT

Candidate SNP	Sequence of probe 5'→3'
rs17192586	AGATTCTTAGGAGTAAC[G/A]GCTGACATTCACCATATT
rs2015599	AATGAATTCTAACTCACT[G/A]CAAGTACTACAGTGTTCCT
rs2038479	ACTGCTATTTTCATTTTAT[C/A]GATGGAATACTTTGAAG
rs2038480	CAAGCGTGTGTTAAGAATA[A/T]GTATATAAAATGTGTTTT
rs214060	AGACAACCGGCAGCTCTAA[C/T]GAAAATATTGGAGACACT
rs2735816	TTTGCCCTGCACTGAGCA[G/C]AGAGCATCTGAAATGTGGA
rs3214051	CGGGGGTGGTGACAAG[G/A]ACTAAAGGGTAAGAATTTA
rs3804749	GCTGCAGGCTGCAAACAGG[C/T]GAAACAGGAAGAGAGA
rs4148450	AACAGGGAACTTGACATC[C/T]GCCCAGACCATCAGTCAAT
rs55905547	AATCTCAGTGTTGTGGGCC[A/G]TAGCGTCCTCACCACA
rs6771416	AACTTCCAGAGACAGCTA[G/A]ATGGGGCAGTGAGTCCAGT
rs7618405	CTTTTGGGAGGCCAC[C/A]ATGAGTTAGCACTCTTTTCT

The complementary strands were annealed using a standard protocol (http://www.piercenet.com/files/TR0045-Anneal-oligos.pdf), consisting of denaturation of the complementary strands to remove any secondary structure and hybridisation of the strands. Annealing occurs most efficiently when the temperature is slowly decreased after denaturation. Complementary oligonucleotides of 100 µM were combined at an equal molar ratio to a final concentration of 1 µM in Buffer EB [QIAGEN]. To anneal the strands, the sample was incubated in a thermocyler [PTC-225 Peltier Thermal Cycler, MJ Research] using the following programme: 5 min at 95°C; 1 min at 94–25°C, with a decrease of 1°C per cycle (70 cycles); hold at 4°C. Double-stranded probes were quantitated in triplicates using a NanoDrop spectrophotometer [ND-1000, Labtech], and analysed in the Nucleic Acid module, DNA-50 mode using the software ND-1000 v3.5.2.

Binding reaction. Gel mobility shift assays were performed with the LightShift Chemiluminescent EMSA Kit [all components, Thermo Fisher Scientific] according to the manufacturer's instructions. Each 20 µl binding reaction contained 1x binding buffer, 75 ng/µl poly(dI/dC), 2.5% glycerol, 0.05% NP-40, 87.5 mM KCl and 6.25 mM $MgCl_2$. For each reaction, 2.5 µl of freshly prepared nuclear protein was used. Biotin-labelled DNA containing the candidate SNP of interest was added in a final amount of 20 fmol. For competition assays, unlabelled probes were added in final amounts of 2 or 4 pmol, representing a 100- or 200-fold molar excess over the labelled probes, respectively (**Table 2-8**). An overview of the incubation times is provided in **Table 2-8**.

Electrophoresis. The binding reaction was then subjected to gel electrophoresis on a native polyacrylamide gel [Novex 6% DNA Retardation Gel, Invitrogen] according to the manufacturer's protocol. The wells of the gel were flushed, and the gel was pre-electrophoresed for 30 min in ice-cold

0.5x Novex TBE Running Buffer [Invitrogen], applying 100 V using the XCell SureLock Electrophoresis Mini-Cell [Invitrogen]. Then, the wells were flushed and loaded with 19 µl of each sample. Samples were electrophoresed for 75 min at 100 V.

Electrophoretic transfer to nylon membrane. After electrophoresis, the gel was extracted from the cassette and carefully transferred onto 0.8 mm blotting paper [Whatman] or filter paper [Mini Trans-Blot, Bio-Rad]. The nylon membrane [Biodyne B, Thermo Fisher Scientific] was soaked in 0.5x Novex TBE Running Buffer [Invitrogen] for at least 10 min. Both blotting/filter paper and blotting pads were briefly soaked prior to use. The blot module was assembled according to the manufacturer's protocol. In brief, the order of assembly was as followed: cathode core, blotting pad, filter paper, gel, nylon membrane, filter paper and blotting pad. The transfer was performed for 1 hr at 30 V (360–270 mA) using a Mini Trans-Blot Cell [Bio-Rad]. Immediately after transfer, the membrane was exposed for 45–60 sec to UV-light (254 nm) [Stratalinker UV Crosslinker 2400, Stratagene], applying 120 mJ/cm^2 in the auto cross-link mode, to cross-link the transferred DNA to the nylon membrane.

Detection of biotin-labelled DNA by chemiluminescence. The biotin-labelled DNA was detected using the Chemiluminescent Nucleic Acid Detection Module [all components, Thermo Fisher Scientific]. The Blocking Buffer and 4x Wash Buffer were gently warmed to 37–50°C in a water bath until all particulate was dissolved. After UV-cross-linking, the membrane was immediately blocked by submerging in 20 ml of Blocking Buffer. The membrane was incubated for 15 min with gentle shaking. The buffer was decanted from the membrane and replaced with 20 ml of Blocking Buffer supplemented with 66.7 µl of Stabilized Streptavidin-Horseradish Peroxidase Conjugate (1:300 dilution). The membrane was incubated for 15 min with gentle shaking, subsequently transferred to a new container and washed four times for 5 min each in 20 ml of 1x Wash Buffer with gentle shaking. The membrane was transferred to a new container and 30 ml of ice-cold Substrate Equilibration Buffer was added. The membrane was incubated for 5 min with gentle shaking. Next, the Chemiluminescent Substrate Working Solution was prepared by adding 5 ml of Luminol/Enhancer Solution to 5 ml of Stable Peroxide Solution, protected from light. The membrane was removed from the Substrate Equilibration Buffer, excess buffer removed, and then placed onto a clean sheet of plastic wrap. The Substrate Working Solution was poured onto the membrane so that it completely covered the membrane. Then, the membrane was incubated for 5 min, protected from light. After incubation, the membrane was removed from excess buffer and covered with plastic foil. Finally, the membrane was placed in a film cassette and exposed to X-ray film [CL-XPosure Film, Thermo Fisher Scientific] for 5–10 min. The film was developed using an X-ray film processor [Compact X4, Xograph] according to manufacturer's instructions.

Quantification of signal density. The blots were quantified by measuring the signal density of the probes competed with the respective unspecific competitor using the software ImageJ v1.45 (http://rsbweb.nih.gov/ij/). The mean density ratio represents the ratio of the measured density of the stronger and the weaker band (0.15 x 1.50 rectangular area centred on each band).

Supershift. For supershift experiments, the antibody was added to the reaction mix at the end, prior to incubation. **Table 2-8** gives an overview of the conditions used in the experiments.

Table 2-8. Overview of the experimental setup for EMSA and supershift experiments. Changes in experimental parameters were based on optimisation experiments.

Candidate SNP	Incubation of binding reaction	Additional agent	Molar excess of competitor	Antibody for supershift experiments
rs342293	120 min at RT	n/a	200-fold	4 µl EVI1 [sc-8707 X, Santa Cruz Biotechnology (SCB)]
	60 min at RT	n/a	200-fold	2 µl GATA1 [ab28839, Abcam] 4 µl RUNX1 [sc-28679 X, SCB]
chr1:145,507,646 (5'-UTR *RBM8A*)	45 min at RT	n/a	100-fold	2 µl EVI1 [sc-8707 X, SCB]
chr1:145,507,765 (intron *RBM8A*)	120 min at RT	0.1 mM EDTA	100-fold	2 µl MZF1 [sc-46179 X, sc-66991 X, SCB] 2 µl RBPJ [sc-28713 X, SCB]
All other (Table 2-7)	45 min at RT	n/a	100- or 200-fold (Figure 4-11)	n/a

2.12. Expression QTL analysis

Data sets. Published gene expression profiling and genotypic data sets were obtained from different sources depending on the cell type studied. Details regarding experimental protocols and data processing can be found in the respective references (**Table 2-9**).

Table 2-9. Genotyping and gene expression platforms used for eQTL analyses. Even though different versions of Illumina platforms were used, the probe for *PIK3CG* was the same across all chips (probe-ID: ILMN_1770433).

Cell type/ tissue	Genotyping	Gene expression profiling	Reference
Platelets	Applied Biosystems TaqMan	Illumina HumanWG-6 v2	Sivapalaratnam et al., *in preparation*
Macrophages	Illumina Human 1.2M-Custom / Illumina Human 670-Quad-Custom	Illumina HumanRef-8 v3	Rotival et al., 2011
Monocytes			
LCLs	Illumina Human 1M-Duo	Illumina HumanHT-12 v3	Nica et al., 2011
Adipose			
Skin			

Analysis. Expression QTL analyses with rs342293 and its proxy SNPs were performed with the software Genevar (Gene Expression Variation) using a window of ±1 Mb centred on the SNP (Yang et al., 2010). The strength of the relationship between alleles and gene expression intensities was estimated using Spearman's rank correlation and reported as nominal *P*-values.

2.13. Whole-genome gene expression profiling of *Pik3cg*$^{-/-}$ mice

Ethics statement. The study had ethical approval from the NHS Cambridgeshire Research Ethics Committee. The care and use of all mice in this study was carried out in accordance with the UK Home Office regulations under the Animals (Scientific Procedures) Act 1986.

***Pik3cg* knockout mice.** *Pik3cg* knockout mice were obtained from sources described in Sasaki et al., 2000, backcrossed onto the C57BL/6J Jax genetic background for eight generations (B6J;129-Pik3cg^{tm1Pngr}) and then maintained as a closed colony by intercrossing from within the colony (C57BL/6J Jax contribution: 99.6%). PCR genotyping was performed with the following primer pairs: 5'-TCA GGC TCG GAG ATT AGG TA, 5'-GCC CAA TCG GTG GTA GAA CT (wild type); 5'-GGA CAC GGC TTT GAT TAC AAT C, 5'-GGG GTG GGA TTA GAT AAA TG (mutant), as described in Sasaki et al., 2000.

Blood collection. Approximately 0.5 ml of whole blood from three *Pik3cg*$^{-/-}$ and three C57BL/6J Jax wild type mice (all females, age: 13–16 weeks, Mouse Breeders Diet [Lab Diets 5021-3]) was collected from terminally anesthetised mice via the retro-orbital sinus.

RNA extraction. Total RNA was extracted using the Mouse RiboPure-Blood RNA Isolation Kit [Ambion] according to the manufacturer's protocol. Total RNA was quantitated and quality-checked using a NanoDrop spectrophotometer [ND-1000, Labtech]. **Table 2-10 A** gives an overview of the RNA extraction yield.

Gene expression profiling. After extraction, 500 ng of total RNA was transformed into biotinylated cRNA using the TotalPrep RNA Amplification Kit [Ambion] according to the manufacturer's protocol. The protocol comprised three main steps: (1) reverse transcription of total RNA to synthesise full-length, first-strand cDNA with an oligo(dT) primer bearing a T7 promoter; (2) second-strand synthesis to convert the single-stranded cDNA into a double-stranded DNA template for *in vitro* transcription, followed by cDNA purification; and (3) *in vitro* transcription with T7 RNA Polymerase and biotin-UTP to generate multiple copies of biotinylated antisense RNA (cRNA) from the double-stranded cDNA templates. Following cRNA purification, the amplified and labelled cRNA was directly used for array hybridisation. **Table 2-10 B** gives an overview of the yield and quality of biotin-labelled cRNA.

Table 2-10. Assessment of quantity and quality of total RNA and biotin-labelled cRNA.

Mouse	(A) RNA extraction	(B) Labelled cRNA preparation			
	Yield	*Repl.*	*Yield*	*260/280*	*260/230*
Pik3cg$^{-/-}$-A	20.55 µg	1	16.43 µg	2.04	2.02
		2	16.00 µg	2.07	2.05
Pik3cg$^{-/-}$-B	17.73 µg	1	15.41 µg	2.07	1.99
		2	21.51 µg	2.06	2.03
Pik3cg$^{-/-}$-C	20.63 µg	1	18.38 µg	2.08	2.05
		2	16.42 µg	2.08	2.04
WT-A	11.65 µg	1	13.81 µg	2.07	1.96
		2	15.90 µg	2.05	1.99
WT-B	25.08 µg	1	13.07 µg	2.06	2.00
		2	17.39 µg	2.05	1.86
WT-C	13.10 µg	1	13.38 µg	2.05	1.97
		2	15.35 µg	2.05	2.02

For each sample, 750 ng of biotinylated cRNA was hybridised to Illumina MouseWG-6 v2 Expression BeadChips. The BeadChip contains 45,281 unique probes that target the NCBI Reference Sequence (RefSeq) database v22 (ftp://ftp.ncbi.nih.gov/refseq/release/), Mouse Exonic Evidence-Based Oligonucleotide (MEEBO) set (http://www.arrays.ucsf.edu/archive/meebo.html) and the exemplar protein-coding sequences described in the RIKEN FANTOM2 database (http://fantom2.gsc.riken.jp/).

After hybridisation, arrays were washed, detected and scanned on a BeadArray Reader [Illumina] according to the manufacturer's protocol.

Data processing. On the raw expression data, background subtraction, variance-stabilising transformation and quantile normalisation were performed across all samples with the R package lumi (Du et al., 2008). Technical replicates were averaged and the differentially expressed transcripts between wild type and knockout mice were identified by calculating the $\log_2$-fold changes of the averaged expression values. *P*-values were calculated by 1-way analysis of variance (ANOVA). All analyses were carried out in the R/Bioconductor environment.

Gene Ontology. Gene Ontology term enrichment analysis was performed using the web-based tool AmiGO v1.7 (Carbon et al., 2009) with the following parameters: gene expression fold-change cut-off: ±1.5; background: MGI; *P*-value cut-off: $1x10^{-5}$; minimum number of gene products: 10.

Data availability. The whole-genome gene expression data sets of *Pik3cg*$^{-/-}$ and wild type mice are available online in the GEO database under accession number GSE26111.

2.14. Protein-protein interaction network

Of the 220 differentially expressed genes between *Pik3cg*$^{-/-}$ and wild type mice (**Section 2.13**), 191 orthologous human genes were retrieved using BioMart (http://www.ensembl.org/biomart/martview/), and their respective proteins using UniProt (http://www.uniprot.org/). These 'core' proteins were used as primary seeds to develop the protein-protein interaction network. First-order interactors of core proteins were determined using Reactome v36 (http://www.reactome.org/). Only clustered non-redundant first-level interactions between human proteins that were connected to the largest connected component were considered. Based on the HaemAtlas data (Watkins et al., 2009), interactors that are not expressed in MKs ($P>0.01$) were excluded. Further details about the approach are described in Gieger et al., 2011.

2.15. Exome sequencing of individuals with TAR syndrome

Ethics statement. Informed consent was obtained from all study subjects with approval from the ethics committees of the following institutions: University Hospital Bristol (MREC/00/6/72), Universitair

Ziekenhuis Leuven (ML-3580), University of Cambridge (REC 10/H0304/66, REC 10/H0304/65), INSERM (RBM 1-14) and Charité Universitätsmedizin Berlin (EA2/170/05).

Exome sequencing and analysis. The 'baits' to capture the complete human exonic sequence were designed using annotations from the GENCODE Consortium, comprising a total of 740,000 exons in 79,000 transcripts from a highly redundant set of 34,642 genes, and covering 39.3 Mb of genomic sequence (Coffey et al., 2011). Exonic sequence was captured and enriched using the SureSelect Human All Exon Kit [Agilent Technologies] and sequenced on the Illumina GAII platform. Sequence analysis was performed as described in Albers et al., 2011. Variants that did not overlap the targeted regions ±25 bp were filtered and not considered for further analyses. The functional consequence of SNPs and small indels were predicted using the Ensembl Variation API (http://www.ensembl.org/info/docs/api/variation/). Each variant was annotated for presence in databases of genetic variation. Sequence variants with allele frequencies of up to 5% were considered, as inferred from variation data from dbSNP v131 (http://www.ncbi.nlm.nih.gov/SNP/), the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) and 354 exomes from the CoLaus Cohort (Firmann et al., 2008). Target enrichment for exome sequencing was performed at the Wellcome Trust Sanger Institute by the sequencing core group.

Data availability. Sequencing data are available online in the European Genome-phenome Archive (EGA) under accession number EGAD00001000018.

2.16. Sanger sequencing of the *RBM8A* locus

Primers for capillary sequencing were designed using Primer3 as described in **Section 2.9**, and are reported in **Table 2-11**. PCR products were amplified from genomic DNA using the ThermoStart Taq DNA Polymerase Kit [Thermo Scientific], and cleaned up using the ExoSAP-IT PCR Product Clean-Up Kit [USB]. Capillary sequencing was performed by Source Bioscience.

Table 2-11. Primer pairs used for Sanger sequencing of the *RBM8A* locus. Genomic coordinates were based on the human reference genome, build hg19.

Forward primer sequence	Reverse primer sequence	Genomic position	Description
CACGCCAGCCTCTGAGTT	CCCTAATCTCAAACCACTTCCT	chr1:145,501,836–145,502,420	upstream reg. element
GGACGAGCAGGAGACAGATG	CGCACCTGGCCTAAAAATTC	chr1:145,502,151–145,502,744	upstream reg. element
CAAAGCACACCCTGCACA	CACCTCCTGGGTTCAGGTAA	chr1:145,502,469–145,502,949	upstream reg. element

Forward primer sequence	Reverse primer sequence	Genomic position	Description
GGCCAGCCTGGGTAGTATAA	CCAGTCTGGGGGACAAGAG	chr1:145,506,316–145,506,883	promoter
TGCACCACTGCACTCTTAGC	TTTAGGCAGCGTGGTGTATG	chr1:145,506,650–145,507,248	promoter
GTCTCCCGGGTTCAACTG	TCTAAATCCCTCCCTCTGCAC	chr1:145,506,920–145,507,559	promoter
GCCCAGCTAATCAGCTTCC	TCCCTCTGCACGGTAAAAAC	chr1:145,506,991–145,507,549	promoter
TTTCCCAGTTTGGGATGAAG	GGGCGGAATCTCTAATCCAC	chr1:145,507,301–145,507,871	genic
GCCGGGCCTCACTGTTAAT	TCAGTTTGTGAATGCTCTCTGG	chr1:145,507,354–145,508,033	genic
GCCGCGGTTAAGAGGAAG	TTGTGAATGCTCTCTGGAACC	chr1:145,507,395–145,508,028	genic
ATGGCCACAGAAACACTTCC	CACCGCCTCCAGTCTTAGTG	chr1:145,507,474–145,507,924	genic
AGTTAGCCTTTGATTGGTCAGC	ACCCGTAGCTCCTGCCCTA	chr1:145,507,474–145,508,174	genic
ATGGCCACAGAAACACTTCC	TCCTCCTTTCTCCCATTGTTC	chr1:145,507,474–145,508,174	genic
ATGGCCACAGAAACACTTCC	CCACAGACACGGATACCTCA	chr1:145,507,474–145,508,324	genic
CGGGTCTTGGGTGGATTA	CCACAGACACGGATACCTCA	chr1:145,507,842–145,508,324	genic
GGGTCTTGGGTGGATTAGAGA	TTTAAGCAGGCTCACAGGAA	chr1:145,507,843–145,508,430	genic
GGGTTCCAGAGAGCATTCAC	GATATCCTGTTCGCCTGTCG	chr1:145,508,007–145,508,606	genic
CCTAGTAGGGCAGGAGCTACG	CCAACCACAGCAAACACAGA	chr1:145,508,105–145,508,692	genic
CGCAGTAGGAATGGGTTCAG	CCTGGGCTTCCTTGTATGTT	chr1:145,508,355–145,508,958	genic
GGCCAAGAGCAAAGTTGAAA	CCCAGTCCTATTTGTCCAAGG	chr1:145,508,691–145,509,284	genic
TTGTCAGACACGCCAAAGAG	CAATGATCCATACAGCCTTGC	chr1:145,508,736–145,509,438	genic
TGGGTGAAGGGAATACGAAC	ATGGTGGCATGTGCCTGTA	chr1:145,509,079–145,509,626	genic
GTGTTACCCAGGGTGGATTG	CATGCCTTTAGACAGCTGGA	chr1:145,509,508–145,509,946	genic
GGGAGGGACTTCAGTTAGCA	CCTGTTGCCTCTAGCATCATT	chr1:145,510,103–145,510,687	genic
TGATAGAAATATGAAGCCACCAAG	AAGGATGAATTGGGAGGAGAC	chr1:145,510,458–145,511,028	genic
AAGAGGCAGCAGAAGGTGAA	CAGCCCAATAGCATTTGGAA	chr1:145,510,814–145,511,453	genic
GGCTTGAATATGATGCTGAACA	GCCTGATCGTAACTCCAAACA	chr1:145,511,127–145,511,721	genic

2.17. Genotyping of the 5'-UTR and intronic SNPs at the *RBM8A* locus

The *RBM8A* 5'-UTR and intronic SNPs were genotyped in 7,504 individuals from the Cambridge BioResource with custom TaqMan SNP Genotyping Assays [Applied Biosystems] according to the manufacturer's protocols. All genotyping data were scored twice by different operators.

2.18. Sequencing of megakaryocyte RNA

Megakaryocyte RNA was prepared and sequenced as described in Albers et al., 2011. Reads were aligned to the *Homo sapiens* high-coverage assembly (build hg19) using the software GSNAP v2011-03-28 (Wu & Nacu, 2010). Read trimming was disabled and up to five mismatches were allowed. Newly

identified splicing sites had to be at most 100 kb apart. Aligned reads were visualised using the Integrative Genomics Viewer (Robinson et al., 2011).

2.19. *RBM8A* promoter activity by luciferase reporter assay

Co-transfection experiments in different cell lines, i.e. EAHY926, HEK296, MC3T3, CHRF-288-11 and DAMI, were performed with pEGFP vectors [Clontech], and *RBM8A* reporter plasmids (wild type or with the 5'-UTR/intronic SNP) were constructed from pGL3-Basic luciferase vectors [Promega]. The *RBM8A* promoter region, starting at -303 nt upstream of the transcription start site and including exon 1 and the first 142 nt of intron 1, was cloned 5' to the luciferase gene. For each co-transfection assay, cells were transfected using Lipofectamine [Life Technologies] with 2 μg of pEGFP and 4 μg of *RBM8A*-pGL3 plasmid for HEK293, EAHY926 and MC3T3 cells. DAMI and CHRF-288-11 cells were transfected using the Amaxa electroporation system [Lonza] according to the manufacturer's instructions. Luciferase activity was determined as described in Freson et al., 2007. Each plasmid was assayed in triplicates in six separate transfection experiments. Firefly luciferase activity was normalised to EGFP expression. Statistical analysis was performed using the software InStat v3.01 [GraphPad].

2.20. Y14 protein expression analysis in platelet extracts

Blood (20 ml) anticoagulated with 3.8% trisodium citrate was centrifuged at 200xg to obtain platelet-rich plasma (PRP). The platelet pellet was obtained by centrifugation at 700xg after addition of 0.1 volume of ACD buffer at pH=4.5 (2.5% trisodium citrate, 1.5% citric acid and 2.0% D-glucose). The platelet pellet was lysed in ice-cold lysis buffer (1.0% Igepal CA-630 [Sigma-Aldrich], 1 mM EDTA, 2 mM DTE and 1x EDTA-free Protease Inhibitor [Complete Mini, Roche] per 50 ml of PBS) and subjected to three freeze-thaw cycles. Lysates were cleared of insoluble debris by centrifugation for 10 min at 4°C and 16,100xg. Protein fractions were mixed with 5% SDS reducing sample buffer, separated by SDS-PAGE and transferred to Hybond ECL-nitro-cellulose membranes [GE Healthcare]. The blots were blocked for 1 hr at RT in Tris-buffered saline with Tween-20 supplemented with 5% non-fat dry milk. The blots were first incubated with primary antibody overnight at 4°C, and then with horseradish peroxidase (HRP)-conjugated secondary antibody for 2 hr at RT. The following primary antibodies were used: rabbit polyclonal antibody against Y14 (Q-24), mouse monoclonal antibody against Y14 (4C4) [Santa Cruz Biotechnology], mouse monoclonal antibody against Gsα (Freson et al., 2001) and mouse monoclonal antibody against β-actin [A5441, Sigma-Alrich]. Both Y14 antibodies

were tested for their specificity using recombinant Y14-GST purified by sepharose beads as described in Freson et al., 2001. Signal was detected with the ECL Western Blotting Substrate [Thermo Fisher Scientific] according to the manufacturer's protocol. Densitometry analysis was carried out using the software ImageJ64.