SUPPLEMENTARY METHODS

Identification of tissue-specific CRMs

A step-wise *in silico* strategy was used to identify tissue-specific CRMs based on the following steps: 1) we first identified tissue-specific genes that are highly expressed in the liver based on statistical analysis of micro-array expression data of normal human tissues[17-18]; 2) we then extracted the corresponding promoter sequences from public databases (NCBI36/hg18 genome assembly); 3) using the TRANSFAC® database we subsequently mapped these TFBS elements to these promoters. We then identifified the tissue-specific CRM using a multidimensional scaling (MDS)/ differential distance matrix (DDM) approach[12]; 4) Finally, the cross-species conservation was taken into consideration to identify highly conserved *HS-CRMs.*

To identify highly expressed tissue-specific genes, we used a high-density gene expression database of 18,927 unique genes based on microarrays that were derived from 158 normal human samples from 19 different organs of 30 different individuals[18]. This was used to identify a set of most highly expressed (i.e. 'over-expressed') genes in the liver compared to any of the other tissues. A two-tailed t-test was used for each pairwise comparison. Conversely, a set of 'under-expressed' genes was identified, corresponding to those genes that exhibited the lowest expression in the liver compared to any of the other tissues. This analysis resulted in a set of 59 liver-specific over-expressed genes and a set of equal size under-expressed genes.

Next, the RefSeq IDs lists of these 'over-expressed' and 'under-expressed' heart or liver-specific genes were used to extract the corresponding promoter sequences upstream the reported transcription start sites (TSSs) by up to 0.8 kb (NCBI36/hg18 genome assembly), using the transcription start location data stored in the refGene table of the UCSC Genome Browser (http://genome.ucsc.edu) database. This resulted in two sets of tissue-specific

promoter sequences corresponding to promoters of 'over-expressed' or 'under-expressed' genes. In order to make a non-redundant set of representative promoter sequences, the promoter sequences were filtered using 'uclust' (http://www.drive5.com/usearch/). The rationale for chosing promoters up to 0.8 kb upstream of the TSS is supported by *in silico* analysis and *in vitro* experiments. The basal promoter and nearby upstream regulatory elements are typically found within a 0.8 kb region upstream of the TSS. Indeed, most known TFBS in the TRANSFAC® (TRANScription FACtor) database (http://www.biobase-international.com/product/transcription-factor-binding-sites) have preferred locations between -300 and +50 bp relative to the TSS[22]. Moreover, luciferase-based transfection assays in four human cultured cell types showed that the large majority (>90%) of DNA fragments containing regions -550 to +50 relative to the TSS were transcriptionally active[23], which is consistent with the TRANSFAC datasets. Finally, we previously showed that the outcome of the DDM-MDS analyses is very similar, regardless as to whether a 800 bp or 1,500 bp upstream promoter region is taken into consideration.

We then mapped the TFBS to these promoters using the TRANSFAC® database. TRANSFAC® is a manually curated database of eukaryotic transcription factors, their genomic binding sites and DNA binding profiles which is used to predict potential transcription factor binding sites. Based on its broad compilation of binding sites, positional weight matrices are derived which can be used with either the Match™ of fimo tool to search DNA sequences for predicted transcription factor binding sites. Subsequently, we used the DDM/MDS method, described in detail elsewhere[12], on the TFBS datasets obtained from the 'over-expressed' versus 'under-expressed' heart or liver-specific genes. The computer source code is deposited in a public repository http://www.dmbr.ugent.be/prx/bioit2-public/TFdiff/TFdiff.tar.gz. This method allowed us to simultaneously identify TFBS that are over-represented in addition to TFBS that tend to cluster together ('co-occurrence') in those promoters of highly expressed liver-specific genes.

In the final step, the genomic context of the tissue-specific over-expressed genes was searched for cross-species conserved regions for the TFBS associated with high tissue-specific gene expression. For that purpose, we downloaded the sequences of all conserved sequence elements in the NCBI36/hg18 genome assembly based on the information stored in the phastConsElements44way table of the UCSC Genome Browser (http://genome.ucsc.edu) database. The predicted conserved sequence elements are assigned a log-odds score equal to its log probability under the conserved model minus its log probability under the non-conserved model. This allows to restrict the search for putative *CRMs* that coincide with the most conserved sequence elements. The conserved sequence elements were scanned for TFBS associated with high tissue-specific expression using the Match^{TM} program or fimo application[24]. Using internally developed Perl scripts, this led to the identification of highly conserved *CRMs* containing clusters of TFBS associated with high tissue-specific expression.

Determination of vector titers

AAV vectors were produced as previously described (16). Titers were determined using TaqMan® probes and primers specific for the hFIX cDNA or the polyadenylation signal, as described[20]. Briefly, The hFIX cDNA forward and reverse primers were: 5'-AGGGATATCGACTTGCAGAAAA-3' and 5'-GTGAGCTTAGAAGTTTGTGAAACAG-3', respectively. The probe used was 5'-FAM-AGTCCTGTGAACCAGCAGTGCCATTTC-TAMRA-3'. The pA forward and reverse primers were 5'-GCCTTCTAGTTGCCAGCCAT-3' and 5'-GGCACCTTCCAGGGTCAAG-3', respectively. The probe was 5'-FAM-TGTTTGCCCCTCCCCCGTGC-TAMRA-3. Briefly, reactions were performed in TaqMan® Universal PCR Master Mix, on an ABI 7500 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA). Known copy numbers (10^2–10^7) of the respective vector plasmids used to generate the corresponding AAV vectors, carrying the appropriate cDNAs, were

used to generate standard curves. To determine titer of the vector used for the non-human primate study, q-PCR using SYBR® Green and bovine growth hormone poly-A (BGHpolyA) primers were used. Forward: 5'-GCCTTCTAGTTGCCAGCCAT-3', reverse: 5'-GGCACCTTCCAGGGTC-AAG-3'. To generate standard curves, known copy numbers of the corresponding vector plasmids were used. Titers typically fell within the normal range of 2-5x10^{12} vector genomes (vg)/ml.

Animal experiments

All animal procedures were approved by the institutional animal ethics committees. FIX-deficient hemophilia B mice were kindly provided by Dr. I. Verma & Dr. L. Wang, The Salk Institute for Biological Studies, USA & Dr. Kay, Stanford University)[21]. Experiments were conducted on 6–7 kg male captive-bred cynomolgus macaques purchased from BioPrim, Baziège, France. The Institutional Animal Care and Use Committee of the Région des Pays de la Loire (University of Angers, France) approved the protocol. Macaques were pre-screened for pre-existing anti-AAV9-neutralizing antibodies and only those with no detectable antibodies were employed in the study. One animal received an immunosuppressive regime consisting of oral administration of cyclosporin A (25 mg/kg) and intravenous administration of rituximab (MabThera®, Roche) (20 mg/kg) along with 4 mg/kg of the antihistamine drug dexchlorphéniramine (Polaramine®, Msd) to prevent allergic reactions against Rituximab.

Collection of samples from non-human primates

Blood samples were collected under ketamine-induced anesthesia (10 mg/kg). Samples that were hemolyzed were discarded and not considered in the analysis of coagulation, hematological and chemical parameters. For surgical liver biopsies, anesthesia was induced with ketamine in combination with diazepam and was maintained using an inhalational mixture of isoflurane and oxygen. Analgesia was performed with morphine, and meloxicam

was administered the following three to five days to avoid animal discomfort. A complete monthly check-up (blood and clinical) was performed on each animal following rAAV administration. Organs were collected following euthanasia, performed by intravenous injection of pentobarbital sodium. For surgical liver biopsies, anesthesia was induced with ketamine in combination with diazepam and was maintained using an inhalational mixture of isoflurane and oxygen. Analgesia was performed with morphine, and meloxicam was administered the following three to five days to avoid animal discomfort. A complete monthly check-up (blood and clinical) was performed on each animal following rAAV administration. Organs were collected following euthanasia, performed by intravenous injection of pentobarbital sodium.

Cell phenotyping

Frequency of B and T cells in peripheral blood was determined by flow cytometry. PBMC were isolated using Ficoll gradient and stained with anti-CD20 and anti-CD3 antibodies to determine circulating B and T lymphocytes percentages respectively. After cell incubation with antibodies, stainings were acquired using a LSR II flow cytometer (BD Bioscience) and analyzed with BD FACSDiva version 6.1.3 software (BD Biosciences).

Monitoring of hFIX antigen, anti-AAV9 capsid IgG, anti-hFIX IgG and Bethesda assay in non-human primates

An ELISA using a hFIX-specific coating antibody (Haemotologic Technologies, Essex Junction, VT), was used to determine hFIX antigen levels in monkey plasma, as previously described[13]. To measure anti-hFIX antibody levels, a capture assay was used, as previously described[13]. Briefly, ELISA plates were coated with hFIX protein and test samples were added. An HRP-conjugated goat anti-monkey antibody was used to detect hFIX specific IgG.

To measure the inhibitory antibodies against the hFIX transgene product, a modified Bethesda assay was used[13]. Briefly, citrated test plasma was heat inactivated for 1 hour at 56°C to eliminate endogenous monkey FIX. Subsequently, incubation with human plasma for 2 hours at 37°C took place. An activated partial thromboplastin time (aPPT) assay was used to measure the residual hFIX activity. One Bethesda Unit (BU) is the reciprocal of the dilution of test plasma at which 50% of hFIX activity is inhibited. Plasma levels of anti-AAV9 antibodies were determined according to a previously described capture assay[13] which was slightly adjusted. Briefly, 1.3 x 10^9 AAV9 vector particles/ml were used to coat the ELISA plates, after which plates were blocked. HRP-conjugated goat anti-monkey antibody was used for detection. A standard curve was generated using serial dilutions of purified monkey IgG.

Statistics

P-values were calculated using Student t-test. Data were analyzed using Microsoft Excel Statistics package. Values shown in the figures were expressed as the mean + s.e.m. or s.d. as mentioned in the legends. A value of $P \leq 0.05$ was set as the level of statistical significance. The different data sets were evaluated for normality using a Kolmogorov-Smirnov test included in SPSS 21.0 statistic software (SPSS Inc., USA) when the data set was <5 and Lilliefors test for normality when data set was ≥5.

Supplementary References

22. Mariño-Ramírez, L., Spouge, J.L., Kanga, G.C. and Landsman, D. (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res.*, **32**, 949–958.

23. Trinklein, N.D., Aldred, S.J.F., Saldanha, A.J. and Myers, R.M. (2003) Identification and functional analysis of human transcriptional promoters. *Genome Res.*, **13**, 308–312.

24. Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinforma. Oxf. Engl.*, **27**, 1017–1018.