nature biotechnology

Massively parallel functional dissection of mammalian enhancers *in vivo*

Rupali P Patwardhan[1,8], Joseph B Hiatt[1,8], Daniela M Witten[2], Mee J Kim[3], Robin P Smith[3], Dalit May[4], Choli Lee[1], Jennifer M Andrie[1], Su-In Lee[1,5], Gregory M Cooper[6], Nadav Ahituv[3], Len A Pennacchio[4,7] & Jay Shendure[1]

The functional consequences of genetic variation in mammalian regulatory elements are poorly understood. We report the *in vivo* dissection of three mammalian enhancers at single-nucleotide resolution through a massively parallel reporter assay. For each enhancer, we synthesized a library of >100,000 mutant haplotypes with 2–3% divergence from the wild-type sequence. Each haplotype was linked to a unique sequence tag embedded within a transcriptional cassette. We introduced each enhancer library into mouse liver and measured the relative activities of individual haplotypes *en masse* by sequencing the transcribed tags. Linear regression analysis yielded highly reproducible estimates of the effect of every possible single-nucleotide change on enhancer activity. The functional consequence of most mutations was modest, with ~22% affecting activity by >1.2-fold and ~3% by >2-fold. Several, but not all, positions with higher effects showed evidence for purifying selection, or co-localized with known liver-associated transcription factor binding sites, demonstrating the value of empirical high-resolution functional analysis.

Massively parallel sequencing has accelerated the cataloging of *cis*-regulatory elements in mammalian genomes[1–3]. Although diverse methods exist to predict the functional consequences of genomic variants that alter protein sequence, it remains challenging to estimate the functional effects of variation in *cis*-regulatory elements[4]. Furthermore, understanding the architecture and internal grammar of *cis*-regulatory elements is essential for advancing our comprehension of the mechanistic basis for regulatory activity, for enabling the *de novo* design of synthetic regulatory elements, and for predicting the functional and phenotypic consequences of genetic variation within noncoding DNA. Recent studies have highlighted the importance of regulatory variants. For example, disease-associated variants frequently coincide with regulatory regions and, in particular, with enhancers[2,5–8].

We previously reported a method called 'synthetic saturation mutagenesis'[9] in which programmable microarrays were used to synthesize variants of several regulatory elements (core promoters), each in *cis* with a downstream tag sequence. The population of core promoter variants was subjected to a cell-free *in vitro* assay, after which sequencing of the transcribed tags was performed to quantify the relative activity of specific core promoter variants. Although successful, several aspects of this approach limit its broader application and scalability: (i) when each regulatory element variant is synthesized as a separate array feature, the overall cost of synthesis remains high; (ii) the separate synthesis of individual variants also limits how many combinations of mutations can be simultaneously programmed; (iii) the maximum length of array-synthesized oligonucleotides is currently 200–300 bp, whereas mammalian enhancers can be 1 kb or longer; (iv) access to array-derived oligonucleotide libraries remains restricted to a few groups; and (v) the cell-free, *in vitro* assay that we used poorly captures biological context.

To overcome these limitations and facilitate the high-resolution dissection of mammalian enhancers, we developed an improved method, termed massively parallel functional dissection (MPFD) (**Fig. 1**). We then used MPFD to assess the extent to which all possible single-nucleotide variants (SNVs) affect the activity of three mammalian enhancers that are active in the liver, designated here ALDOB (hg19:chr9:104195570-104195828)[10–12], ECR11 (hg19:chr2:169939082-169939701)[13,14] and LTV1 (mm9:chr7:29161443-29161744). We demonstrate here that MPFD generates highly complex libraries of enhancer haplotypes where all possible single-nucleotide substitutions are present on many individual haplotypes, and that MPFD yields reproducible estimates of the effect of these substitutions on enhancer activity. We find that most substitutions result in modest effects on enhancer activity, and that most co-occurring substitutions affect activity independently of one another. We also find that evolutionary constraint and predicted transcription factor binding sites are often but not always concordant with estimates of MPFD effect size, emphasizing the importance of direct experimental characterization.

RESULTS

To apply the MPFD method (**Fig. 1**) to the three enhancers of interest, each enhancer was synthetically constructed by polymerase cycling

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Biostatistics, University of Washington, Seattle, Washington, USA. [3]Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA. [4]Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, California, USA. [5]Department of Computer Science, University of Washington, Seattle, Washington, USA. [6]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [7]Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to J.S. (shendure@uw.edu) or L.A.P. (lapennacchio@lbl.gov) or N.A. (nadav.ahituv@ucsf.edu).

Received 15 September 2011; accepted 23 January 2012; published online 26 February 2012; doi:10.1038/nbt.2136

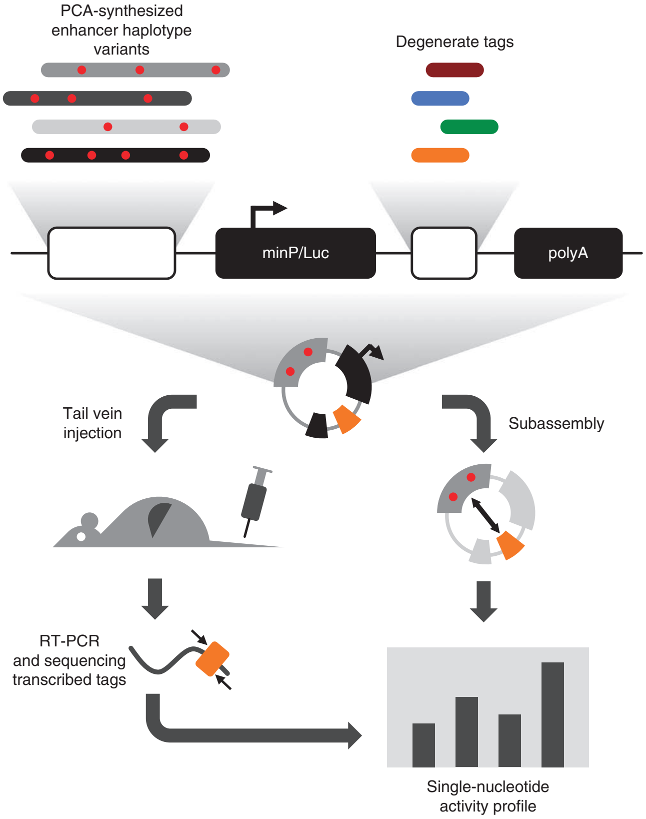
PCA-synthesized enhancer haplotype variants
Degenerate tags
minP/Luc
polyA
Tail vein injection
Subassembly
RT-PCR and sequencing transcribed tags
Single-nucleotide activity profile

Figure 1 Overview of MPFD. We used doped oligonucleotide synthesis and polymerase cycling assembly (PCA) to generate a highly complex library of enhancer haplotypes for each enhancer studied. On average, each enhancer haplotype diverged from wild type by ~2–3% (red circles represent mutations). These mutant enhancers, along with 20-bp degenerate tags, were cloned into an expression vector (pGL4.23) containing a minimal promoter driving transcription of luciferase (minP/Luc). We performed 'subassembly' on each library to determine the full sequence of each enhancer haplotype and to identify the 20-bp tag to which each haplotype was cloned in *cis*. Each library was then introduced into two mice through hydrodynamic tail vein injection, livers were harvested after 24 h and sequencing was performed to quantify abundance of transcribed 20-bp tags. These data were used to estimate the effect of each possible mutation on transcriptional activation.

assembly using overlapping oligonucleotides (~90 bp) that contain a programmed level of degeneracy. At each position, 97% of molecules were expected to be synthesized correctly with 1% doping of each possible single-nucleotide substitution (Online Methods). Therefore, each synthetic enhancer molecule contained, on average, three mutations per 100 bp, randomly distributed along its length. The population of molecules was inherently complex, both with respect to representation of all possible SNVs of the wild-type enhancer as well as myriad unique combinations. Because nearly all synthetic enhancers contained multiple substitutions, they are referred to here as 'enhancer haplotypes'.

Next, a library for assessing the activity of each enhancer haplotype was created by cloning the synthetic enhancers into a plasmid (Promega pGL4.23), which contains a minimal promoter upstream of the luciferase gene. In order to uniquely tag each enhancer haplotype, we cloned an oligonucleotide containing a 20-bp, fully degenerate subsequence to a separate site in the 3′ untranslated region (UTR) of the luciferase gene. The sequences of specific 20-bp tags cloned in *cis* with specific enhancer haplotypes were determined by massively parallel sequencing. As the enhancer haplotypes were highly related sequences with lengths that exceeded the maximum read-length of the Illumina platform, we used tag-guided subassembly[15] to enable full-length, high-accuracy sequencing of individual enhancer haplotypes in association with their downstream tags. Each resulting library included >100,000 fully sequenced enhancer haplotypes, with nearly all containing multiple substitutions, and each associated with one or more unique tags.

The library was then subjected to what was effectively a massively parallel *in vivo* reporter assay. For the experiments described here, we used the hydrodynamic tail vein assay[13,16] to assess *in vivo* enhancer activity in the mouse liver. Mice were euthanized 24 h after injection, at which time total RNA was extracted from each liver, followed by RT-PCR and massively parallel sequencing of cDNA from transcribed tags.

MPFD of three enhancers

We studied three mammalian enhancers identified by diverse methods (**Supplementary Fig. 1**). ALDOB (259 bp) is a human intronic enhancer of the aldose B gene[10–12]. ECR11 (620 bp) is a human enhancer located in an intron of dehydrogenase/reductase SDR family member 9 (*DHRS9*)[13]. LTV1 (302 bp) is a candidate mouse enhancer located on the 3′ side of zinc-finger protein 36 (*Zfp36*) (**Supplementary Fig. 2a**,**b**). The activity of each wild-type enhancer was confirmed using a conventional hydrodynamic tail vein injection assay, in which luciferase activity in liver tissue was measured 24 h after injection (**Supplementary Fig. 2c**).

We applied MPFD to systematically dissect the functional consequences of all possible SNVs in these three enhancers (**Fig. 1**). Sequencing with subassembly confirmed that the resulting libraries were complex, with a total of 641,135 distinct haplotypes associated with 1,186,696 tag sequences (**Table 1**). The observed number of mutations per haplotype approximated expectations, with ~2–3 substitutions per 100 bp (**Supplementary Fig. 3**) and were well distributed (**Supplementary Fig. 4**). All possible substitution variants of each enhancer were represented in ≥42 uniquely tagged haplotypes. On average, each position was disrupted on ~4,000 distinct enhancer haplotypes. Furthermore, all possible pairs of positions were disrupted in ≥1 haplotype with the exception of a single pair of positions in LTV1.

Table 1 Enhancer haplotype library characteristics

Library	Number of haplotypes	Number of tags	Percent of possible substitutions in at least one haplotype	Percent of possible pairs of positions in at least one haplotype	Per-base mutation rate per haplotype (mean ± s.d.)
ALDOB	378,450	406,071	100 (777 of 777)	100 (33,411 of 33,411)	0.021 ± 0.010
ECR11	105,795	105,832	100 (1,860 of 1,860)	100 (191,890 of 191,890)	0.023 ± 0.006
LTV1 rep. 1	119,950	403,869	100 (906 of 906)	99.99 (45,449 of 45,451)	0.031 ± 0.010
LTV1 rep. 2	105,188	270,924	100 (906 of 906)	99.99 (45,449 of 45,451)	0.031 ± 0.010

For each library of enhancer haplotypes, we list the number of distinct haplotypes, the number of tags with which those distinct haplotypes are associated in *cis*, the percentage of possible single-nucleotide substitutions that are present in at least one haplotype, the percentage of possible pairs of positions where both positions contain mutations together in at least one haplotype and the per-base mutation rate in each library.

We introduced each library (one each for ALDOB and ECR11, and two independently constructed libraries for LTV1) into two mice by hydrodynamic tail vein injection (**Supplementary Fig. 2d**). Total RNA from each mouse liver was split into several aliquots (ALDOB: N = 39; ECR11: N = 69; LTV1-1: N = 10; LTV1-2: N = 10), with each aliquot separately subjected to RT-PCR with primers flanking the 20-bp tag located in the 3′ UTR of the luciferase transcriptional cassette, and then to massively parallel sequencing on an Illumina GAIIx. Because target RNA was very scarce relative to cellular RNA, a modest number of target RNA molecules contributed to each RT-PCR, leading to a complexity bottleneck. In other words, within each sequencing library, all reads corresponding to any single tag appeared to have been derived from amplification of a single RNA molecule. We therefore used the number of RNA aliquots in which a particular tag was observed, and not the total number of reads associated with a tag, as a measure of the relative transcriptional activity of its associated enhancer haplotype.

For each position in each enhancer, we constructed a linear model to assess the extent to which the presence of a mutation at that position is predictive of a change in the number of RNA aliquots in which an enhancer haplotype was observed, which is effectively a proxy for its effect on transcriptional activation, that is, 'effect size' (Online Methods). Specifically, we use the term 'effect size' to describe the $\log_2$-fold change in the predicted transcriptional activity, as measured

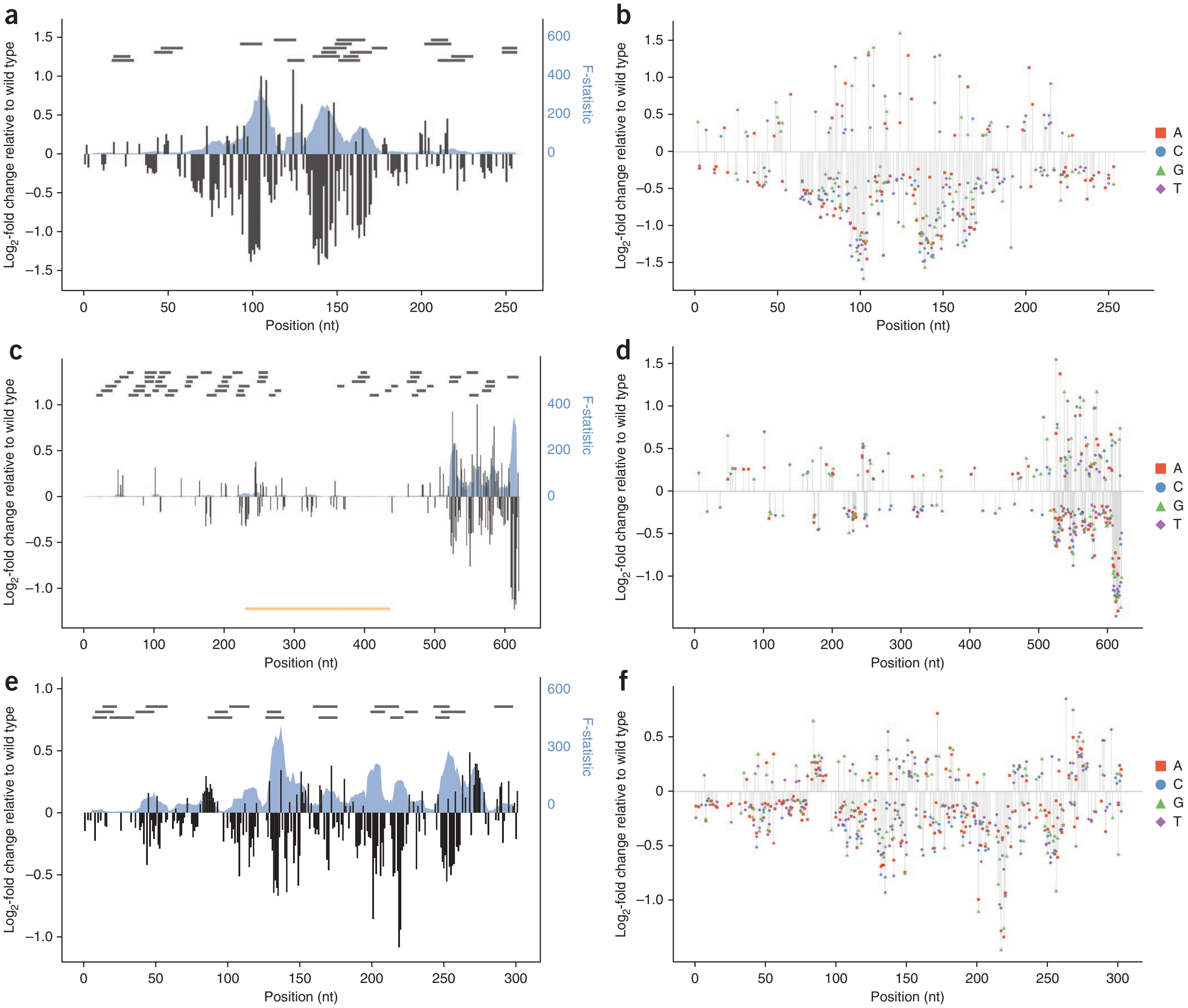

Figure 2 Effect size on transcriptional activity of all possible substitution mutations in three mammalian enhancers. (**a**–**f**) Estimated effect size of mutation at each position based on coefficients from univariate (gray columns, left axis) and trivariate (A:red, C:blue, G:green, T:purple) models are shown for ALDOB (**a** and **b**, respectively), ECR11 (**c** and **d**, respectively) and LTV1 (**e** and **f**, respectively). Effect sizes were estimated by taking the $\log_2$ of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide (total number of aliquots sequenced per library: ALDOB: 39; ECR11: 69; LTV1 set 1: 10; LTV1 set 2: 10). Effect sizes are shown only for positions where model coefficients had associated P-values ≤ 0.01. We also used multiple linear regression with sets of ten adjacent positions as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (blue shadow, right axis) for ALDOB (**a**), ECR11 (**c**) and LTV1 (**e**). The locations of TFBS predictions using the MATCH web server (with restriction to TFs present in liver) are shown as horizontal gray bars at the top of the plot in **a**, **c** and **e**. The location of a partial LINE element in ECR11 is shown as an orange bar at the bottom of **c**.

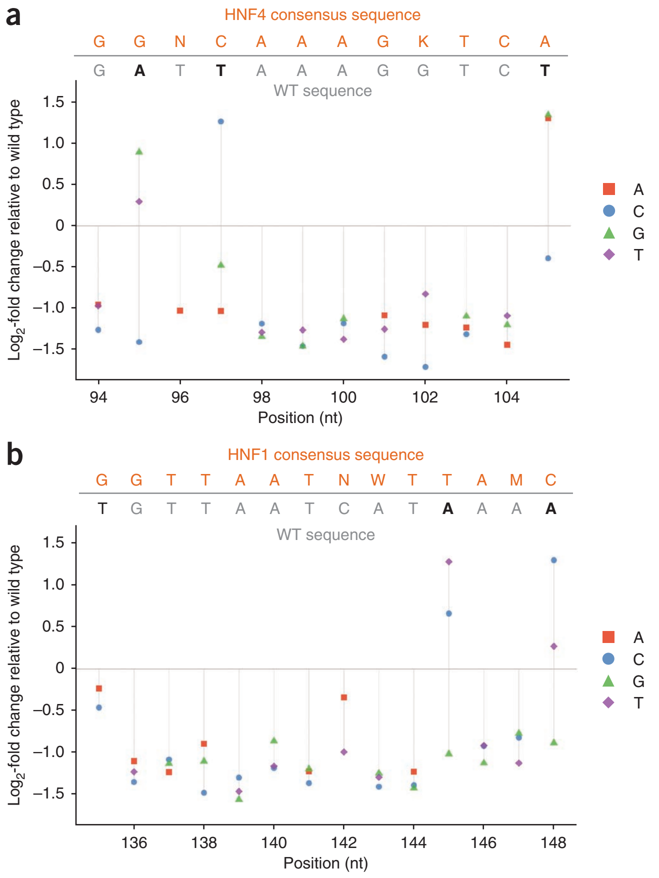

Figure 3 Profiles of mutation effect size in TFBSs. (**a**,**b**) For a predicted HNF4 site (positions 94–105) (**a**) and a predicted HNF1 site (positions 135–148) (**b**) in ALDOB, the effect size for each possible substitution, with the consensus TF binding sequence (orange) and the enhancer sequence (gray for consensus, black for nonconsensus) is plotted. Nonconsensus positions where rescue is observed after mutating to consensus are shown in boldface. HNF4 binding to the ALDOB enhancer region in human liver has been previously demonstrated[22], whereas *in vivo* occupancy data for HNF1 at this region is not yet available.

by the number of RNA aliquots in which a tag-associated haplotype appeared, relative to the wild type. We first sought to assess reproducibility, so we calculated effect sizes separately for the two independently constructed LTV1 libraries (combining data from the two mice subjected to each of these libraries). For ALDOB and ECR11, we calculated effect sizes separately on the data from each mouse. For these two types of biological replicates, the effect sizes were highly correlated ($r = 0.96$ for LTV1, $r = 0.93$ for ALDOB, $r = 0.96$ for ECR11). Because reproducibility was high and to increase resolving power, we performed all subsequent analyses after combining data across mice for each enhancer haplotype library (data for one of the two LTV1 replicate libraries is shown in **Supplementary Fig. 5**).

We next recalculated effect sizes in two ways (**Fig. 2**). First, as for the reproducibility analysis, we constructed separate linear models for each position where mutational status was encoded as a single binary variable representing whether an enhancer haplotype was wild type or mutant at that position (**Fig. 2a**,**c**,**e** and **Supplementary Table 1a**). Second, we constructed separate multiple linear regression models for each position with three variables, each corresponding to a particular nucleotide substitution at that position (**Fig. 2b**,**d**,**f** and **Supplementary Table 1b**). For each enhancer, we also constructed a multiple linear regression model incorporating all positions. These models were also significantly predictive ($P < 0.01$) (**Supplementary Note** and **Supplementary Table 2**), and yielded effect-size profiles similar to models constructed independently for each position (**Supplementary Fig. 6**). As the coefficients from models constructed independently for each position are more naturally interpreted as position-specific effects, we used these models for subsequent analyses.

To provide further validation, we also performed site-directed mutagenesis to individually introduce the six mutations in ALDOB that were predicted to have among the largest effect sizes (three increasing activity and three decreasing activity), and tested these individually using the hydrodynamic tail vein luciferase assay (**Supplementary Fig. 7**). Observed luciferase fold-changes were highly correlated with effect-size predictions from the models ($R = 0.985$).

Co-localization of high-impact positions and known TFBSs

Across each enhancer, the effect-size profiles exhibited spatial structure—that is, a clustering of positions with larger effect sizes. Positions separated by less than ~6 nucleotides had significantly correlated effect sizes ($P < 0.01$) (**Supplementary Fig. 8**). To further explore this, we performed multiple linear regression using mutational status at ten adjacent positions (that is, a binary variable for wild-type or mutant) at a time (Online Methods). These models remained predictive of transcriptional activity in a spatially resolved pattern (**Fig. 2a**,**c**,**e**). We suspected that these clusters of correlated positions might represent transcription factor binding sites (TFBSs). Indeed, when we predict TFBSs[17] (**Fig. 2a**,**c**,**e** and **Supplementary Table 3**), we observe striking overlap between predicted binding sites and clusters of highly predictive positions (**Fig. 2a**,**c**,**e**). For example, a predicted binding site for HNF4 in the ALDOB enhancer (bases 94–105) coincides with a highly predictive localized model (**Fig. 2a**). Furthermore, all mutations in this region had negative effects on activity, with the notable exception of mutations that increased identity with the consensus HNF4 binding site, which were activating (e.g., 95A→G and 105T→A) (**Fig. 3a**). The same pattern was observed for other predicted sites as well, for example, a predicted HNF1 binding site at bases 135–148 in ALDOB (**Fig. 3b**). Notably, independent experiments have established that these two transcription factors drive this element *in vivo*[12]. The spatial patterns may also reveal or refine broader features of activity—for example, the boundaries of functional elements. For example, in ECR11, computational prediction yielded a large number of predicted liver-specific TFBSs in the proximal 300 bases[13], but we observed that the highest impact SNVs were largely confined to the distal 160 bases (**Fig. 2d** and **Supplementary Fig. 9**).

Relationship between evolutionary and functional constraint

Evolutionary constraint in noncoding, regulatory DNA has frequently served as a proxy for functional constraint[18–20]. However, recent studies have shown that many enhancers are evolving rapidly and that mammalian genomes contain large numbers of evolutionarily young, sometimes species-specific, enhancers[21,22]. All three enhancers studied here are grossly conserved between human and mouse (**Supplementary Fig. 1**). We therefore investigated the relationship between functional constraint and evolutionary constraint at single-nucleotide resolution. For two of three enhancers, linear models, constructed to assess whether evolutionary constraint (that is, Genomic Evolutionary Rate Profiling (GERP)[23]) was predictive of functional constraint (that is, the absolute value of univariate model coefficients that we obtained), were significantly predictive with

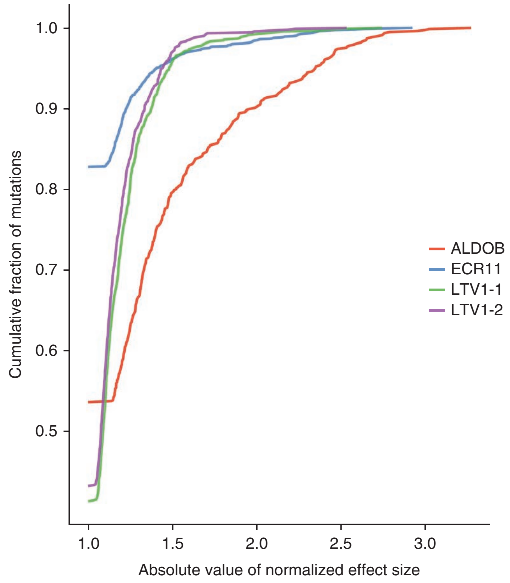

Figure 4 Distribution of effect sizes for all possible substitution mutations in three mammalian enhancers. For the three enhancers studied (two replicate libraries for LTV1), the cumulative fraction of substitutions possessing a given effect size is expressed as the absolute value of the effect size of a given substitution. For example, across the three enhancers, between ~80% and ~95% of substitutions influence transcriptional activity by less than a factor of 1.5.

modest explanatory power (ALDOB: $R^2 = 0.1232$, $P = 6.31e\text{-}9$; LTV1: $R^2 = 0.03911$, $P = 5.47e\text{-}4$). For both enhancers, positions with the highest functional effect sizes were significantly associated with elevated evolutionary constraint scores ($P < 0.01$) (**Supplementary Fig. 10**). However, not all positions with high GERP scores (≥4) had functional effect sizes in the top quartile for each enhancer (ALDOB: 33 of 61, 54%; ECR11: 5 of 25, 20%; LTV1: 0 positions with GERP≥4). These positions might have functions unrelated to the enhancer activity assayed here or might be of greater functional relevance in other contexts, for example, other tissues or developmental time points. On the other hand, a small set of highly functional positions, for example, most nucleotides within the distal-most C/EBP motif in ECR11, have low GERP scores, consistent with lineage or species-specific activity.

Effect-size spectrum of single-nucleotide variants

A substantial proportion of polymorphisms and new mutations in mammalian genomes are single-nucleotide substitutions[24]. However, the functional dissection of regulatory elements has historically relied on introducing nested or scanning deletions, limiting the extent to which they inform the interpretation of naturally occurring variation. Our results provided an opportunity to examine the distribution of effect sizes of SNVs in mammalian enhancers on the magnitude of transcriptional activation (**Fig. 4**). Notably, we observed that the majority of SNVs result in only a modest change in transcription relative to the wild-type enhancer. Overall, <25% of the mutations alter transcriptional activity by >1.2-fold. Furthermore, only a few mutations, mostly in ALDOB, altered activity by a factor of >2. These results suggest that these enhancers are highly robust to the vast majority of potential SNVs. Further application of this method will be needed to assess whether this is a general property of mammalian enhancers.

Perhaps as expected, the majority of functionally important mutations decreased activity (70% or 850/1,211). In general, only one substitution at a given position was activating, for example, substitutions that render a motif more like the consensus sequence (**Fig. 3**). However, we observed some notable exceptions, including positions 83–93 and 272–278 in LTV1, where all or almost all substitutions were activating, consistent with binding of a repressive transcription factor. Positions 83–93 harbor a predicted binding site for NF-1, whereas there are no predicted sites in the immediate vicinity of positions 272–278, highlighting the value of experimental assessment of mutational impact.

Epistatic interactions

Finally, we sought to leverage the fact that our enhancer libraries contain multiple mutations on each haplotype to assess the degree of epistasis, or interaction, between positions in the enhancer. To obtain adequate power, we restricted our analysis to pairs of positions that were both mutated in at least 20 haplotypes. For each pair of positions that passed this cutoff, we built a multiple linear regression model consisting of three binary variables where the first two variables encoded mutation status (wild type or mutant) at each position independently and the third encoded whether both are mutant in a particular haplotype. With a false-discovery rate (FDR) cutoff of 0.05, we observed few pairs with a significant interaction term (ALDOB: 82 of 33,389, 0.25%; ECR11: 199 of 184,206, 0.10%; LTV1: 45 of 43,975, 0.10%), suggesting that the effects of multiple SNVs on the same haplotype are generally additive, or that our study lacked power to identify subtle interactions. Interacting pairs were significantly enriched for proximity (that is, pairs within 10 bp of each other versus pairs further apart, ALDOB: $P < 1e\text{-}4$; ECR11: $P < 1e\text{-}3$; LTV1: $P < 1e\text{-}4$), and we observed several different classes of interacting pairs with respect to the signs of the individual position effects and the sign of the interacting term (**Supplementary Table 4**).

DISCUSSION

We developed a strategy to construct complex libraries of mammalian enhancers that contain all possible single-nucleotide substitutions and hundreds of thousands of distinct haplotypes. This method surpasses its predecessor[9] in terms of cost effectiveness, tunability, applicability to full-length regulatory elements and integration with an *in vivo* assay. We applied this method to empirically measure the distribution of effect sizes of all possible SNVs in three mammalian enhancers in an *in vivo* model. A key finding is that the vast majority of SNVs in these enhancers have highly reproducible yet remarkably modest effects on transcriptional activation. The distribution suggests that enhancers are highly robust to single-nucleotide changes. We also find that most combinations of single-nucleotide changes have additive effects on function. As expected, there is a clear relationship between the magnitude of functional impact and the location of predicted TFBSs, although not all predicted TFBSs are functional, and not all functional motifs are associated with predicted TFBSs. Similarly, evolutionary constraint, although clearly correlated with the magnitude of functional impact, does not predict it well on a nucleotide-by-nucleotide basis.

There remain some limitations of the method. First, although we exploited a mouse tail vein assay to assess function *in vivo*, the regulatory elements are episomal and therefore may not be subject to the same mechanisms governing elements residing on chromosomes. For example, because of the size of the synthetic construct, we were unable to assess the effects of mutations that may influence long-range interactions between regulatory elements. This might be addressed in part by transitioning to a lentiviral system, which would facilitate use in additional tissues and may also enable the application of other assays, for example, ChIP-Seq, to enhancer variant libraries. Furthermore, our results must also be considered specific to the minimal promoter used here until other promoter classes are tested. Second, we have assayed these enhancers in a single tissue and at a single time point. The activity profile of specific positions could well be different in other tissues; this is the long-standing context problem[25]. Third, because of the scarcity of the target transcript relative to total RNA, we observed complexity bottlenecking, limiting the precision of our estimates of the effect size. This can be addressed by optimization of the RNA isolation step, for example, by hybridization-based enrichment. Fourth, we restricted our analysis to enhancer haplotypes containing only substitutions, as this was the dominant form of variation introduced during synthesis. To facilitate simultaneous dissection of the functional consequences of small insertions and deletions (indels), one could use reduced-fidelity oligonucleotide synthesis conditions, or polymerase cycling assembly with oligonucleotides containing programmed indels. Current efforts are directed at implementing these improvements, scaling this method to more enhancers and applying it to other classes of noncoding regulatory elements.

A fundamental goal of modern biology is to understand the human genome at single-nucleotide resolution. Single-nucleotide differences between genomes are causative for, or affect susceptibility to, a host of diseases, and single-nucleotide mutations are a primary source of raw material for evolution. We anticipate that the high-throughput, empirical measurement of the functional impact of single-nucleotide variants in enhancers will substantially facilitate the analysis of noncoding variants in genome-wide association study hits, the study of the mechanistic basis for enhancer activity and the engineering of enhancers with desired properties. Furthermore, with cost-effective, massively parallel methods for functional analysis, it may soon be realistic to empirically measure the functional effects of all possible single-nucleotide changes in all noncoding regulatory elements in the human genome.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession code. Raw sequencing reads have been submitted to the NCBI Short Read Archive under accession number SRA049159.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank R. Qiu and J. Kitzman for advice on experimental strategies, and B. Cohen and D. Pe'er for helpful discussions. This work was supported in part by grants HG003988 from the National Human Genome Research Institute (L.A.P.), US National Institutes of Health (NIH) grant DP5OD009145 (D.M.W.), National Institute of General Medical Sciences (NIGMS) award number GM61390 (N.A.), National Institute of Child Health and Human Development (NICHD) grant number R01HD059862 (N.A.), the Pilot/Feasibility grant from the University of California, San Francisco Liver Center (P30 DK026743) (N.A.), AG039173 from the National Institute on Aging (J.B.H.) and a fellowship from the Achievement Rewards for College Scientists Foundation (J.B.H.). M.J.K. was supported in part by NIH Training grant T32 GM007175 and the Amgen Research Excellence in Bioengineering and Therapeutic Sciences Fellowship. R.P.S. is supported by a CIHR fellowship in the area of hepatology. Parts of the research were conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NICHD, NHGRI or the NIGMS.

AUTHOR CONTRIBUTIONS

R.P.P., J.B.H., N.A., L.A.P. and J.S. conceived of key aspects of the project and planned experiments. R.P.P., M.J.K., R.P.S., D.M., C.L. and J.M.A. performed experiments. R.P.P., J.B.H., D.M.W. and G.M.C. analyzed the data. D.M.W. and S.-I.L. contributed guidance with statistical analyses. R.P.P., J.B.H. and J.S. wrote the manuscript. All authors commented on and revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
2. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
3. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
4. Cooper, G.M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
5. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
6. Noonan, J.P. & McCallion, A.S. Genomics of long-range regulatory elements. *Annu. Rev. Genomics Hum. Genet.* **11**, 1–23 (2010).
7. VanderMeer, J.E. & Ahituv, N. *cis*-regulatory mutations are a genetic cause of human limb malformations. *Dev. Dyn.* **240**, 920–930 (2011).
8. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
9. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
10. Sabourin, J.C. *et al.* An intronic enhancer essential for tissue-specific expression of the aldolase B transgenes. *J. Biol. Chem.* **271**, 3469–3473 (1996).
11. Gregori, C. *et al.* Expression of the rat aldolase B gene: a liver-specific proximal promoter and an intronic activator. *Biochem. Biophys. Res. Commun.* **176**, 722–729 (1991).
12. Gregori, C., Porteu, A., Mitchell, C., Kahn, A. & Pichard, A.L. In vivo functional characterization of the aldolase B gene enhancer. *J. Biol. Chem.* **277**, 28618–28623 (2002).
13. Kim, M.J. *et al.* Functional characterization of liver enhancers that regulate drug-associated transporters. *Clin. Pharmacol. Ther.* **89**, 571–578 (2011).
14. Dang, Q. *et al.* Structure of the hepatic control region of the human apolipoprotein E/C-I gene locus. *J. Biol. Chem.* **270**, 22577–22585 (1995).
15. Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* **7**, 119–122 (2010).
16. Zhang, G., Budker, V. & Wolff, J.A. High levels of foreign gene expression in hepatocytes after tail vein injections of naked plasmid DNA. *Hum. Gene Ther.* **10**, 1735–1737 (1999).
17. Kel, A.E. *et al.* MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**, 3576–3579 (2003).
18. Loots, G.G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
19. Margulies, E.H. *et al.* Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17**, 760–774 (2007).
20. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40**, 158–160 (2008).
21. Blow, M.J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).
22. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
23. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
24. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
25. Botstein, D. & Shortle, D. Strategies and applications of in vitro mutagenesis. *Science* **229**, 1193–1201 (1985).

ONLINE METHODS

Data availability. In addition to raw sequencing reads available in the NCBI SRA, a full list of mutations interrogated for this work, along with the associated effect sizes and *P* values, are provided as **Supplementary Data**.

Construction of enhancer haplotypes from short, doped oligonucleotides using PCA. Sets of overlapping oligonucleotides for each enhancer were designed either by manual inspection (LTV1) or using the program DNAWorks (ALDOB and ECR11). Common flanking sequences were included on either side to allow for amplification of the full-length enhancer haplotypes during PCA. For LTV1, two versions of overlapping oligonucleotides were designed, such that the overlap region in each was different. Oligonucleotides were synthesized by Integrated DNA Technologies (IDT). All positions corresponding to the enhancer region were synthesized using a hand-mix doped at a ratio of 97:1:1:1 (that is, designated base at a frequency of 97%, and every other base at a frequency of 1%). Sequences of all oligonucleotides are listed in **Supplementary Methods**.

For ALDOB as well as ECR11, the full-length haplotypes were assembled in a single step. We used 50 fmol of each oligonucleotide (ALDOB_PCA_OLIGO[1…6] or ECR11_PCA_OLIGO[1…12]) in a 25 µl PCR reaction volume with 1× KapaHiFi Hot Start Ready Mix (Kapa BioSystems), and 0.5× SYBR Green II, with the following cycling conditions: 95 °C for 3 min; followed by 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. Four such reactions were carried out in parallel and then pooled together for each enhancer. The PCR product representing a complex pool of enhancer haplotypes was purified using QIAquick columns (Qiagen). The assembled enhancer haplotypes were then subjected to an additional around of PCR to add 15 bp of vector homology on either side to render them competent for cloning using InFusion (Clontech). We used 20 ng of template in a 25 µl PCR reaction volume with 1× KapaHiFi Hot Start Ready Mix, 0.5× SYBR Green II, and each primer (VH_F and VH_R) at 0.3 µM final concentration. Thermal cycling was done with the following program: 95 °C for 3 min; followed by 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. Sixteen such reactions were carried out in parallel and then pooled together for each enhancer. The PCR product was purified using QIAquick columns (Qiagen).

The two LTV1 designs were assembled separately. For each design, pairs of oligonucleotides, that is, oligonucleotides 1 and 2, oligonucleotides 3 and 4, and oligonucleotides 5 and 6, were each assembled in parallel and the products of the three reactions were then assembled together into the final product in a single reaction. The combinations of primers and oligonucleotides used in each reaction are listed in **Supplementary Methods**. Each 50 µl PCR reaction was prepared on ice with 1× iProof Ready Mix (Bio-Rad), 0.5× SYBR Green II, forward and reverse primers each at 0.5 µM final concentration and 50 fmol of each template oligo. Thermal cycling was done in a MiniOpticon Real-time PCR system (Bio-Rad) with the following program: 98 °C for 30 s, followed by 30 cycles of 98 °C for 10 s, 62 °C for 30 s and 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. PCR products were purified on a QIAquick column (Qiagen). The haplotypes obtained from each of the two LTV1 designs were pooled after the PCA step. Two aliquots were drawn from this pool, and then carried through subsequent steps as two independent samples and were associated with entirely different sets of tags.

Cloning of enhancer haplotypes and the degenerate tag into pGL4.23. For ALDOB and ECR11, we first cloned in the degenerate tag to create a complex library of tagged pGL4.23 plasmids. We then cloned in the enhancer haplotypes into these tagged pGL4.23 plasmids. For LTV1, we first cloned in the enhancer haplotypes and then cloned in the degenerate tag. Details of each cloning step remained the same, irrespective of the order in which they were carried out, and are described below.

Cloning of degenerate tag into pGL4.23 plasmid. The tag oligonucleotide (TAG_OLIGO) was made double-stranded using primer extension in a 50 µl reaction volume with 1× iProof Master Mix, 0.5 µg single-stranded tag oligo, 0.5 µg reverse primer (TAG_EXTEND). The reaction was incubated at 95 °C for 3 min, 61 °C for 10 min and then 72 °C for 5 min. The product was purified using a QIAquick column and eluted in 50 µl EB. It was further subjected to ExoI treatment in 40 µl reaction volume for 1 h at 37 °C to degrade any remaining single-stranded DNA, and purified again using QIAquick columns. The resulting double-stranded tag oligo was then cloned into pGL4.23 at the XbaI site (at 1,799 bp) using standard InFusion (Clontech) protocol. The InFusion reaction was diluted to 100 µl using TE8. We used 1.5 µl of this diluted cloning reaction to transform 50 µl of chemically competent FusionBlue cells (Clontech) using the standard protocol. When the tag was being cloned in first, 16 such transformation reactions were pooled and grown overnight in four 50-ml liquid cultures at 37 °C in a shaking incubator. DNA was extracted using the Invitrogen Charge Switch Mini Prep Kit for ALDOB and ECR11, and the Invitrogen Charge Switch Midi Prep Kit for LTV1.

Cloning enhancer haplotypes into pGL4.23 vector. The enhancer haplotypes were cloned into the EcoRV site (at 42 bp) of the pGL4.23 plasmid, using standard InFusion protocol. We used 1.5 µl of the cloning reaction to transform 50 µl of chemically competent FusionBlue cells using standard protocol. Five transformations reactions were pooled and grown overnight in 50 ml liquid cultures at 37 °C in a shaking incubator. DNA was extracted using the Invitrogen Charge Switch Mini Prep Kit for ALDOB and ECR11, and the Invitrogen Charge Switch Midi Prep Kit for LTV1.

Tail vein injections. Enhancers were injected using methods as previously described[13]. Briefly, each library was injected into mice using the *Trans*IT EE Hydrodynamic Gene Delivery System (Mirus Bio) following the manufacturer's protocol. We injected 10 µg of each library, alongside 2 µg of pGL4.74[*hRluc*/TK] vector to correct for injection efficiency, into the tail vein of CD1 mice (Charles River). After 24 h, mice were euthanized and livers were harvested.

Measurement of luciferase activity. Firefly and renilla luciferase activity were measured on a Synergy 2 Microplate Reader (BioTek Instruments) for each liver using the Dual Luciferase Reporter Assay System (Promega). The firefly luciferase to renilla luciferase ratios were determined and expressed as relative luciferase activity. All mouse work was approved by the UCSF Institutional Animal Care and Use Committee.

Isolation of RNA from mouse livers. Fresh liver tissue was immediately stabilized in RNAlater solution (Ambion). Samples were homogenized in TRIzol reagent (Invitrogen) and RNA was isolated from the samples according to the manufacturer's instructions.

DNase treatment of RNA. To remove any DNA contamination in the RNA extracted from mouse livers, it was subjected to DNaseI treatment using DNA-*free* (Ambion). Each reaction was prepared with 1× DNA-free buffer, 1 µl of rDNaseI enzyme, 10 µg of RNA and RNase-free water to 50 µl. The reactions were incubated at 37 °C for 1 h, with an additional 1 µl of enzyme added midway through the incubation. The reaction was stopped by adding 7 µl of the inactivation reagent and incubating for 2 min at 25 °C with frequent shaking. The reaction was centrifuged in a microcentrifuge at 10,000*g* for 1.5 min, and the supernatant containing RNA was carefully transferred to a fresh tube.

RT-PCR. Aliquots of RNA obtained after DNase treatment were reverse transcribed to cDNA and amplified by PCR using the Qiagen One-Step Kit. The PCR sought to amplify the 20-bp degenerate tag encoded at the 3′ end of the luciferase transcript. The reactions were assembled on ice in a 25 µl total volume with the following reagents: 1× Qiagen One-Step RT-PCR buffer, 400 µM of each dNTP, 0.6 µM of forward primer (BARCODE_PE_F), 0.6 µM of relevant reverse primer (BARCODE_PE_R_ILMN_INDEX[1-8]), 0.5× SYBR Green II and 5 µl (~1 µg) of RNA template. Thermal cycling was done on a Bio-Rad MiniOpticon Real-Time PCR system with the following program: 50 °C for 30 min (reverse transcription), 95 °C for 15 min (inactivation of reverse transcriptase and heat-activation of the DNA polymerase), then 30 cycles of 94 °C for 30 s, 65 °C for 30 s and 72 °C for 30 s. Each reaction was monitored and extracted from the PCR machine when the fluorescence began to plateau. The cDNA products were purified using the QIAquick PCR

doi:10.1038/nbt.2136

Purification Kit (Qiagen) and eluted in 35 µl EB. The primers used for the RT-PCR contained the necessary sequences for compatibility with the Illumina flow-cell. Thus, the cDNA library obtained at the end of this step was ready for sequencing, eliminating the need for a separate sequencing-library construction step. The reverse primer additionally included 6 bp barcodes allowing for several RT-PCR reactions to be pooled into a single lane for sequencing.

Sequencing of RNA-derived tags. The pooled RT-PCR reaction products were sequenced on an Illumina GAIIx using a sequencing primer (BARCODE_SEQ_F) designed to read into the tag sequence. Each run was 36 cycles with an additional 6 cycles to read the indexing barcode (index sequencing primer: BARCODE_PE_R_ILMNINDX[1-8]).

For each aliquot, reads were filtered based on the quality scores for the first 20 bases, which correspond to the degenerate tag. The numbers of occurrences of each tag were counted and tags that were supported by at least ten reads were classified as being 'present' in that aliquot.

Associating tags with enhancer haplotypes. The enhancer haplotypes and tags were situated more than 1,000 bp away from each other on the pGL4.23 plasmid. To bring them adjacent and facilitate the subassembly method, we digested the pGL4.23 plasmids using HindIII, which had two cut sites, one just 3′ of the enhancer, and one just 5′ of the tag, thus resulting in excision of the intervening region. Cut site 1 was already a part of the pGL4.23 backbone. Cut site 2 was engineered in as a part of the tag oligo. The digest was carried out in a 50 µl volume with 1× NEB Buffer 2, 1 µg of plasmid and 1 µl of HindIII Enzyme (New England BioLabs) and incubated at 37 °C for 3 h. The digested plasmid was purified using a QIAquick column.

The digested plasmids were then recircularized using intramolecular ligation, resulting in the tag becoming adjacent to the 3′ end of the enhancer. Ligation was performed using T4 DNA ligase (New England BioLabs) in a 20 µl reaction with 15 ng of template per reaction. The reaction was incubated for 15 min at 25 °C, followed 20 min at 65 °C to inactivate the ligase.

The enhancer and tag region were amplified from recircularized plasmids using PCR with the forward primer targeting the region immediately 5′ of the enhancer (ENHANCER_F for ALDOB and ECR11, and LTV1_F for LTV1) and the reverse primer targeting the region immediately 3′ of the tag (BARCODE_PE_R). The reaction was carried out in a 25 µl volume with 1× KapaHiFi Hot Start Ready Mix (Kapa BioSystems), 0.5× SYBR Green II, 5 µl of the ligation reaction, and each primer at 0.3 µM final concentration. Thermal cycling was done using Bio-Rad MiniOpticon Real-Time PCR system using the following program: 95 °C for 3 min; and then 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each reaction was monitored and removed from the PCR machine when the fluorescence began to plateau. The reactions were then pooled and purified using QIAquick columns.

The amplicons were then subjected to the subassembly protocol as conceptually described[3] with some modifications as follows. The random fragmentation step was carried out using the Nextera Tn5 transposase (EpiCentre) instead of mechanical shearing. The Nextera reaction was purified using MinElute column (Qiagen) and size-selected by PAGE (LTV1: 100+; ECR11:100-300,300+; ALDOB: no size-selection performed). The size-selected fragments were subjected to PCR in a 25 µl reaction volume with 1× KapaHiFi Hot Start Ready Mix (Kapa BioSystems), 0.5× SYBR Green II, 5 µl of the ligation reaction, Nextera Adaptor 1 at 10 nM final concentration, and primers Nextera BP1 and BARCODE_PE_R at 0.3 µM final concentration each. Thermal cycling was carried out using BioRad Mini Opticon System using the following program: 95 °C for 3 min; and then 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each reaction was monitored and removed from the PCR machine when the fluorescence began to plateau. The PCR products were purified using a QIAquick column and then sequenced on either an Illumina GAIIx or a Hi-Seq 2000. Read1 collected 76 bp/101 bp of the enhancer sequence staring at random breakpoints along the enhancer. Read 2 collected the 20-bp tag sequence.

The reads were then grouped by tag. Reads belonging to each group were then aligned to the wild-type enhancer sequence to identify the mutations on the haplotype associated with that tag using a custom analysis framework.

Estimation of effect size of mutation at each position along the enhancer (univariate model). All linear regression analyses were done using the lm() or lsfit() functions available in the R Statistical Package. To quantity the effect of mutation at any given position on the number of aliquots in which an enhancer haplotype was observed, we built a separate linear regression model at every position along the enhancer, with a single predictor representing whether the given position was wild type or mutant. The predictor was thus a binary variable representing presence (1) or absence (0) of a mutation at that position.

$$y_i = \beta_{0j} + \beta_{1j} X_{ij}$$

where, y_i = number of aliquots in which the ith haplotype was observed (referred to as aliquot counts), and $X_{ij} = 1$ if position j was mutant and 0 if position j was wild type in the ith haplotype.

To facilitate comparison between positions and between enhancers, we calculated the effect size of mutation at a position j as

$$\log_2\left(\frac{\beta_{0j} + \beta_{1j}}{\beta_{0j}}\right)$$

The P-value reported by the model for β_{1j} was used to judge whether the effect size was significant.

For LTV1, as a single haplotype was typically associated with multiple tags, we normalized the aliquot counts for a given haplotype by dividing by the number of tags associated with that haplotype. In the case of ALDOB and ECR11, as the enhancer haplotypes were cloned in second, almost all haplotypes were associated with single tags, and thus the aliquot counts for tags were used directly as the aliquot counts of their linked haplotypes.

Estimation of effect size of each specific nucleotide change at each position along the enhancer (trivariate model). To explore whether the estimated effect sizes for each position were being driven by specific nucleotide substitutions, we modified the model just described to include three predictors, each representing one of the three possible nucleotide substitutions at that position. The factors were set up as binary variables representing the presence (1) or absence (0) of the particular change at that position.

$$y_i = \beta_{0j} + \beta_{1j} X_{ij1} + \beta_{2j} X_{ij2} + \beta_{3j} X_{ij3}$$

Effect sizes were then calculated from the coefficients produced by the models as follows (for $k = 1,2,3$):

$$\log_2\left(\frac{\beta_{0j} + \beta_{kj}}{\beta_{0j}}\right)$$

The P-value reported by the model for β_{kj} was used to judge whether the effect of a given nucleotide substitution at a given position was significant.

Spatial structure. To quantify whether nearby positions tend to have similar effect sizes, we calculated the sum of the absolute values of the differences in effect sizes between positions located at a given distance (lag) from each other. In other words, we calculated

$$S(k) = \sum_{j=k+1}^{N} \left| r_j - r_{j-k} \right|,$$

where $k = 1,2,\ldots,20$ denotes the lag, N denotes the length of the enhancer, and r_i is the effect size of position i.

For each value of the lag k, we also calculated $S_{1*}(k),\ldots,S_{1000*}(k)$, each of which measures the sum of the absolute values of the differences in effect sizes between positions at a distance k from each other, after permuting the effect sizes $(r_1,\ldots,r_N)$. We then calculated a P-value associated with each value of the lag k as the fraction of the $S_{1*}(k),\ldots,S_{1000*}(k)$ that was as small or smaller than $S(k)$.

Models to estimate combined predictive power of blocks of adjacent positions. To further characterize the nature of the spatial structure of the

doi:10.1038/nbt.2136

effect sizes and to explore whether certain regions along the enhancer were enriched for positions with larger effect sizes, we focused on blocks of adjacent positions in a 10-bp sliding window along the length of the enhancer. For each window, we built a multiple linear regression model with one predictor for each position within the window. Each predictor was set up as a binary variable denoting the presence (1) or absence (0) of mutation at that position. The response variable y was the number of aliquots in which a given haplotype was seen.

$$y_i = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{i(j+1)} + \cdots + \beta_{10} X_{i(j+9)}$$

The F-statistic from each model was used as a measure of the collective predictive power of positions within each window.

Multiple linear regression models based on the entire haplotype. The multiple linear regression model included one predictor for each position along the enhancer, encoded as a 1 or 0 to indicate presence or absence of a mutation at that position on a given haplotype, and the response variable y represented the number of aliquots in which the haplotype was observed. Here N is the number of positions within a given enhancer.

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_N X_{iN}$$

A *P*-value for the model was calculated by comparing the mean squared error (MSE) of the model to MSEs of 200 models built using randomly shuffled versions of the response variable. A *P*-value for the model was estimated by calculating the fraction of times that the MSE for models built using a shuffled response vector was at least as small as the MSE computed using real data.

We then expanded the model, such that each position was represented by three predictors to indicate which of the three possible nucleotide substitutions was observed at that position.

$$y_i = \beta_0 + \beta_{1j} X_{i1_1} + \beta_{2j} X_{i1_2} + \beta_{3j} X_{i1_3} + \cdots + \beta_{1j} X_{iN_1} + \beta_{2j} X_{iN_2} + \beta_{3j} X_{iN_3}$$

A *P*-value for the model was calculated by repeatedly permuting the outcome vector as described immediately above; however, only 100 permutations were used, due to the high computational burden of constructing this model.

Identification of epistatic interactions (that is, nonadditive effects) among pairs of mutations. For each pair of positions, we built a linear multiple regression model with three predictors: one predictor each to indicate the presence (1) or absence (0) of a mutation at each of the two positions and a third (referred to as the "interaction term") whose value was set to 1 if both positions were mutant on the given haplotype and 0 otherwise. Only pairs of positions that were both mutant on at least twenty haplotypes were considered.

$$y_i = \beta_{0jk} + \beta_{1jk} X_{ij} + \beta_{2jk} X_{ik} + \beta_{3jk} X_{ij} X_{ik}$$

We used the *P*-values for the interaction terms for the resulting models to calculate a FDR for each interaction term (using the p.adjust() function in R, with method = "BH"). Interaction terms with FDR < 0.05 were considered significant and used for downstream analyses of epistatic interactions.

doi:10.1038/nbt.2136