Guo et al. BMC Genomics (2017) 18:394
DOI 10.1186/s12864-017-3785-4

BMC Genomics

RESEARCH ARTICLE Open Access

Transversions have larger regulatory effects than transitions

Cong Guo[1,2], Ian C. McDowell[1,3], Michael Nodzenski[4], Denise M. Scholtens[4], Andrew S. Allen[5,6], William L. Lowe[7] and Timothy E. Reddy[1,6,8*]

Abstract

Background: Transversions (Tv's) are more likely to alter the amino acid sequence of proteins than transitions (Ts's), and local deviations in the Ts:Tv ratio are indicative of evolutionary selection on genes. Whether the two different types of mutations have different effects in non-protein-coding sequences remains unknown. Genetic variants primarily impact gene expression by disrupting the binding of transcription factors (TFs) and other DNA-binding proteins. Because Tv's cause larger changes in the shape of a DNA backbone, we hypothesized that Tv's would have larger impacts on TF binding and gene expression.

Results: Here, we provide multiple lines of evidence demonstrating that Tv's have larger impacts on regulatory DNA including analyses of TF binding motifs and allele-specific TF binding. In these analyses, we observed a depletion of Tv's within TF binding motifs and TF binding sites. Using massively parallel population-scale reporter assays, we also provided empirical evidence that Tv's have larger effects than Ts's on the activity of human gene regulatory elements.

Conclusions: Tv's are more likely to disrupt TF binding, resulting in larger changes in gene expression. Although the observed differences are small, these findings represent a novel, fundamental property of regulatory variation. Understanding the features of functional non-coding variation could be valuable for revealing the genetic underpinnings of complex traits and diseases in future studies.

Keywords: Transitions, Transversions, Massively parallel reporter assay, SNPs, Regulatory variation

Background

There are millions of candidate gene regulatory elements across diverse human cell types, tissues, and environmental conditions (e.g. [1–4]). Genetic variation in those candidate regulatory elements contributes heavily to the variation in gene expression between individuals and, in turn, to the heritability of complex human traits and diseases [5–7]. Determining the specific genetic contributions to both molecular and organismal traits remains a major challenge, however. That challenge persists, in part, because it is difficult to predict the effect that a given variant or set of variants is likely to have on gene regulation. Overcoming that challenge is important for both basic and translational studies of the genetics of gene regulation [8, 9].

A wide variety of studies have now investigated the genetic contributions to human gene expression. Studies of the associations between genotype with gene expression has revealed genetic contributions to nearly every human gene and across diverse cell types and tissues [10]. Meanwhile, studies of the allele-specific binding of transcription factors (TFs) suggest that noncoding variants can alter gene regulation by several mechanisms: disrupting TF binding directly, disrupting complexes of regulatory factors, and disrupting the underlying chromatin state [11–14]. A challenge the above studies face is that genotypes near each other in the genome are highly correlated due predominantly to limited and non-random sites of meiotic recombination across the human genome (i.e. linkage disequilibrium). As one solution to that challenge, investigators have used reporter gene

* Correspondence: tim.reddy@duke.edu
[1]Center for Genomic and Computational Biology, Duke University Medical School, Durham, NC 27710, USA
[6]Department of Biostatistics and Bioinformatics, Duke University Medical School, Durham, NC 27710, USA
Full list of author information is available at the end of the article

expression assays to measure the effects of genetic variation on the activity of regulatory elements across the genome [15, 16]. In a standard reporter gene expression assay, a regulatory element drives expression of a visually observable reporter gene such as a fluorescent or chemiluminescent protein. By assaying regulatory elements with different genotypes, it is possible to identify genetic variants that directly alter the activity of those elements. Recently, high-throughput versions of those assays have been developed to measure the regulatory effects of many genetic variants and mutants at once [17–20]. In such assays, the regulatory elements drive expression of DNA-encoded barcodes that allow for readout with high-throughput sequencing.

While there are many ways to investigate how genetic variants influence gene regulation, performing those studies in the primary cells and tissues that are most relevant to organismal biology remains challenging. For that reason, understanding which variants to prioritize for testing will be highly valuable. More generally, determining which types of mutations are most likely to influence gene regulation will also be important for studying role of regulatory variation in evolution. As a step towards that long-term goal, we focused on testing whether there are effect differences between the two types of genetic mutations, transitions (Ts's) and tranversions (Tv's). Transitions are DNA mutations that maintain the same number of rings in the nucleotide base, specifically exchanging a one-ring pyrimidine with another pyrimidine, or a two-ring purine for another purine. Transversions, in contrast, are mutations that change the nucleotide base from a purine to a pyrimidine or vice versa. It is well known that Ts's are enriched over Tv's in protein-coding regions of the human genome. One of the reasons that Tv's are thought to be depleted in exons is that they are more likely to result in an amino acid substitution. That difference between the rate of Ts's and Tv's is a foundational principle for studies of the molecular basis for evolution [21–23]. In contrast, the different effects that Ts's and Tv's have in the non-coding genome has not been as well studied. One of the major ways the genetic mutations alter regulatory element activity is by influencing the affinity of TFs to the genome [15, 16, 24]. TFs bind DNA based on both sequence and shape [25]. Here, we show that Tv's are more likely than Ts's to alter local DNA structure, TF binding and, in turn, regulatory element activity. To do so, we integrated data from numerous orthogonal studies of the genetic effects on DNA structure, TF binding, and regulatory element activity. While much remains to be understood, our findings enhance understanding of the effects of genetic variation on human gene regulation.

Results

Tv's alter DNA minor groove width and roll more than Ts's

We first hypothesized that Tv's have a greater impact on the shape of DNA than Ts's. We tested that hypothesis using an empirically-based model of the effect of DNA sequence on DNA shape [26] that has been used previously to investigate the shape readout of TFs [27]. We used that model to predict the effect of Tv's and Ts's embedded in the center of 501 bp DNA sequences on four DNA shape parameters: minor groove width (MGW), propeller twist (ProT), helical twist (HelT), and roll (Fig. 1). Transversions had substantially greater effects on minor grove width (2 Å vs 1.3 Å, an increase of 1.5×) and on roll (10.2° vs 4.4°, an increase of 2.3×). In contrast, the Ts's had greater effects than Tv's on HelT and ProT, but the magnitude of the effects was much smaller (1.09× and 1.14×, respectively). Overall, these results indicate that Tv's overall have a greater impact than Ts's on DNA shape, and disproportionately alter the minor groove width and roll of DNA.

Tv's have greater impacts on predicted and experimentally-measured TF binding

We next hypothesized that Tv's also have greater effects on TF binding than Ts's. We first evaluated whether Tv's have a greater predicted effect according to computational and statistical models of the TF:DNA interaction. To do so, we calculated the change in the position weight matrix (PWM) score of every possible single nucleotide mutation in every TF binding motif in the JASPAR database [28]. Briefly, a PWM quantifies the affinity of a TF to each nucleotide in a potential binding site. The center of the TF binding motif is typically more specific than either edge of the motif and, similarly, mutations near the center of the motif typically had greater impacts on the PWM score. Across all motifs, Tv's had a significantly greater effect on TF binding score both in the center of the motif and in the flanking regions (Fig. 2a). The effect was most pronounced at motif positions with moderate nucleotide specificity (i.e. information content [29]), suggesting that degeneracy in TF binding motifs more often accommodate Ts's than Tv's (Fig. 2b). Together, these results indicate that, across all TFs, Tv's are predicted to have a greater impact on TF binding than Ts's.

To test whether the computational predictions are realized in the genome, we next investigated whether there were also greater differences in TF binding between alleles at Tv's than at Ts's. To do so, we analyzed publicly available allele-specific ChIP-seq data for 42 TFs in the fully-sequenced diploid lymphoblastoid cell line (LCL) GM12878, and for CTCF across six LCLs [30]. In both instances, SNPs with evidence of allele-specific TF binding were subtly but significantly enriched for Tv's when

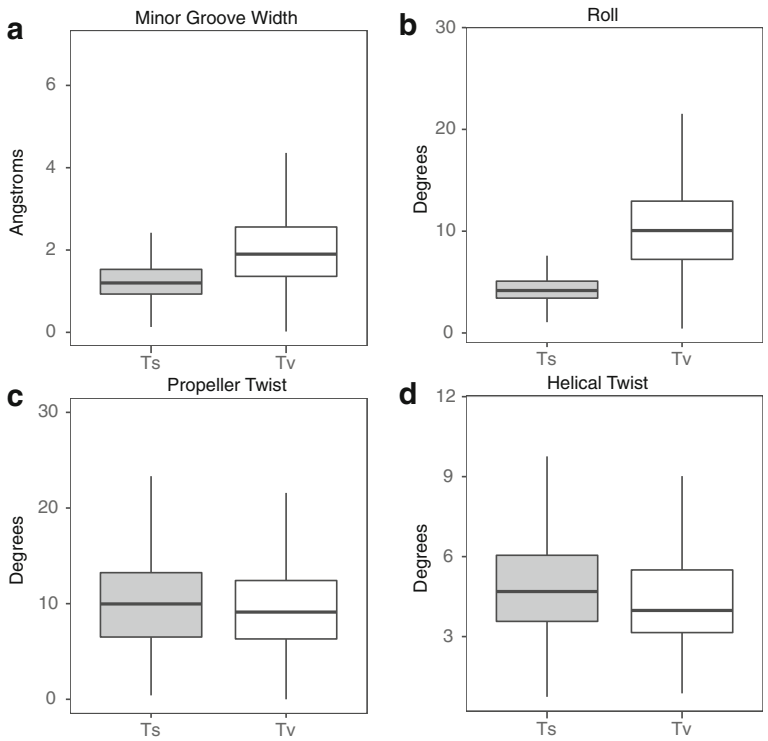

Fig. 1 Tv's have greater effects on DNA shape. Each plot shows the estimated effect of Ts's and Tv's placed in the center of random 500 bp nucleotide sequences on (**a**) minor groove width, (**b**) roll, (**c**), propeller twist, and (**d**) helical twist, as estimated by DNAshapeR. Each plot shows data for all four possible variants of 100,000 random sequences, for a total of 400,000 random sequences analyzed

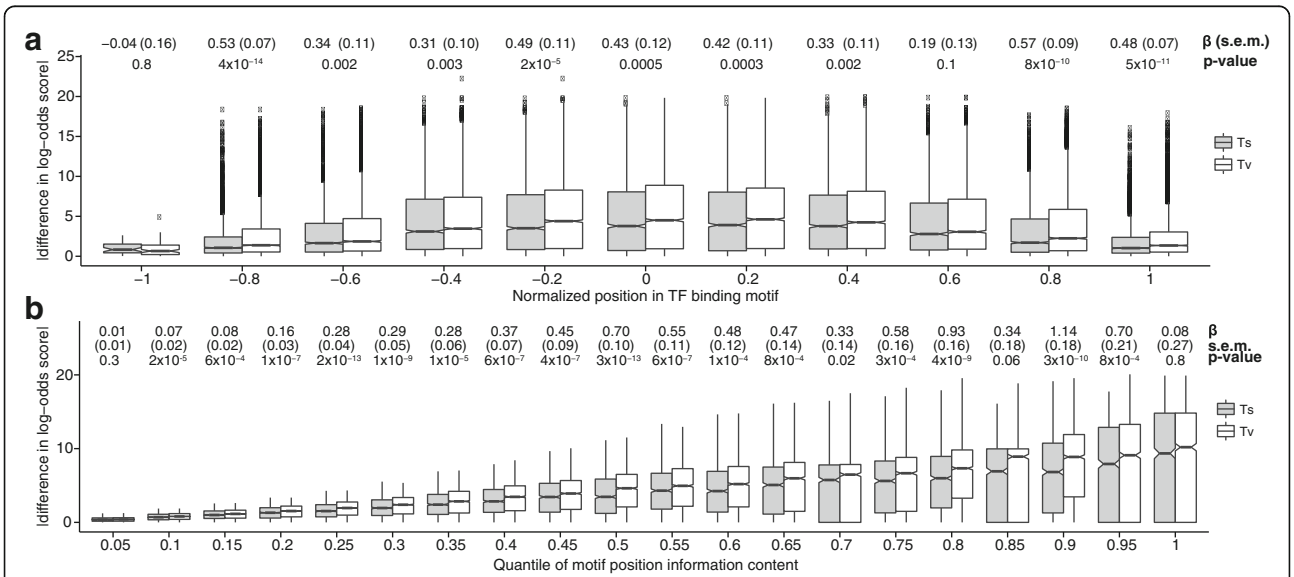

Fig. 2 Tv's are depleted at TF binding sites. **a**, **b** Changes to PWM scores for JASPAR TFs caused by Ts's (*grey boxplots*) and Tv's (*white boxplots*) ordered by normalized position (*above*) and information content (*below*). For each position in each consensus sequence, we calculated the position specific scoring matrix (PSSM) score of having every possible nucleotide at that position and, subsequently, the change in PSSM score when mutating any nucleotide to any other nucleotide at that position. *Black* squares are changes >1.5× above the interquartile range. *P*-values for parameter estimates were calculated using a t test, and are reported without adjustment for multiple hypothesis testing. We tested 11 and 20 hypotheses in Fig. 2a and b, respectively. Applying a Bonferroni correction to a nominal *p*-value threshold of α = 0.05 gives a significance threshold of α = 0.005 and α = 0.0025 for 2A and 2B, respectively

compared to the other SNPs tested (39.33% vs 34.16% for TFs in GM12878 cells, 37.4% vs 34.16% for CTCF in LCLs; Z test $p < 2.2 \times 10^{-16}$, 2.86×10^{-4}, respectively, Fig. 3). These results suggest that Tv's have larger effects on TF binding, resulting in their depletion within TF binding motifs and sites.

Tv's have greater impacts on functional regulatory element activity

Based on our results showing that Tv's have greater effects than Ts's on TF binding, we next hypothesized that Tv's would also have greater effects on regulatory element activity. Because non-coding variation is expected to be a major contribution to human traits and diseases, we focused our analysis on variants on a region on chromosome 3 in which we previously found genetic variants associated with birth weight and fetal adiposity [31–33]. We chose that region as a representative example of a region of the human genome that is associated with a complex human trait or disease. Within that region, we focused specifically on 104 regions that are hypersensitive to digestion by DNaseI (i.e. DNase hypersensitive sites, or DHS's) (Additional file 1: Figure S1, Additional file 2: Table S1). DHS are strongly indicative of TF binding and regulatory element activity, and are also strongly enriched for gene regulatory SNPs associated with human traits [34–36]. By focusing our experiments on DHS at a trait associated locus within a population, we increased our likelihood of capturing expression modulating variants that are relevant to a human phenotype.

To measure the regulatory activity of diverse haplotypes of the 104 DHSs, we used a high-throughput population-scale self-transcribing active regulatory region sequencing (STARR-seq) reporter assay that we call POP-STARR [37]. Briefly, in POP-STARR, candidate regulatory elements from a population of individuals are cloned into the 3′ untranslated region (UTR) of the STARR-seq reporter gene [38]. From that position, each regulatory element control expression of a reporter gene in which it is embedded. For example, once the library is transfected into cells, the regulatory elements with a high level of activity are found frequently in the pool of expressed reporter gene mRNA relative to the regulatory elements with low activity. One can then measure the abundance, and therefore activity, of all regulatory elements in the library by using massively parallel DNA sequencing. Importantly, genotype is observed by DNA sequencing as well. The result of that measurement, therefore, is an allele-specific measure of regulatory element activity across the captured regulatory elements and across the population of from which the regulatory elements were captured.

To test our hypothesis that Tv's have a greater effect on regulatory element activity than Ts's, we use POP-STARR to assay the activity of 104 DHSs captured from the genomes of 760 donors (Fig. 4a). In total, we assayed 1153 unique haplotypes comprised of 942 variants. Of those variants, 634 were Ts's and 308 were Tv's (Additional files 3 and 4: Tables S5-S6). To then test if Tv's alter the activity of regulatory elements more than Ts's, we classified haplotypes by whether they contained a Tv or a Ts relative to a reference haplotype. We then used a multiple linear regression model to test if the presence or absence of a Tv or Ts correlated with changes in regulatory activity between haplotypes. The presence of a Tv

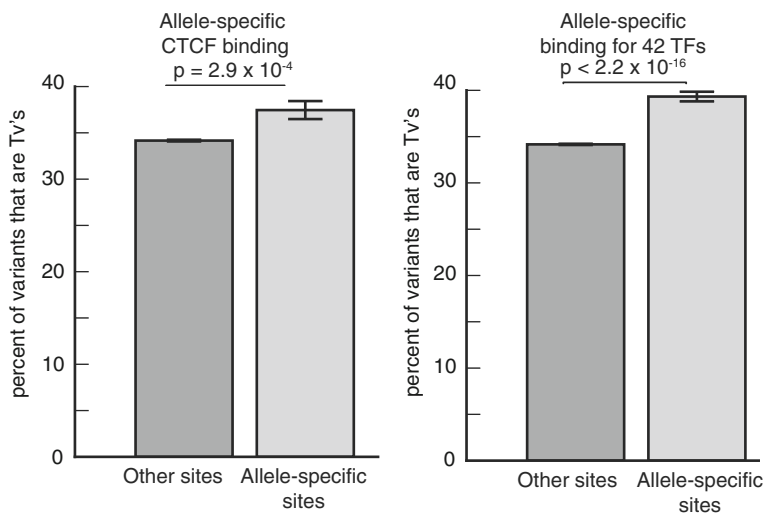

Fig. 3 Tv's are enriched in allele-specific transcription factor binding sites across the genome. The percentage of Tv's in allele-specific CTCF across six LCL lines (*left*), and for allele-specific binding across seven TFs (*right*) in a publically available database [30]. *P*-values and parameter estimates were calculated using a two-tailed Z-test. Error bars show the s.e.m

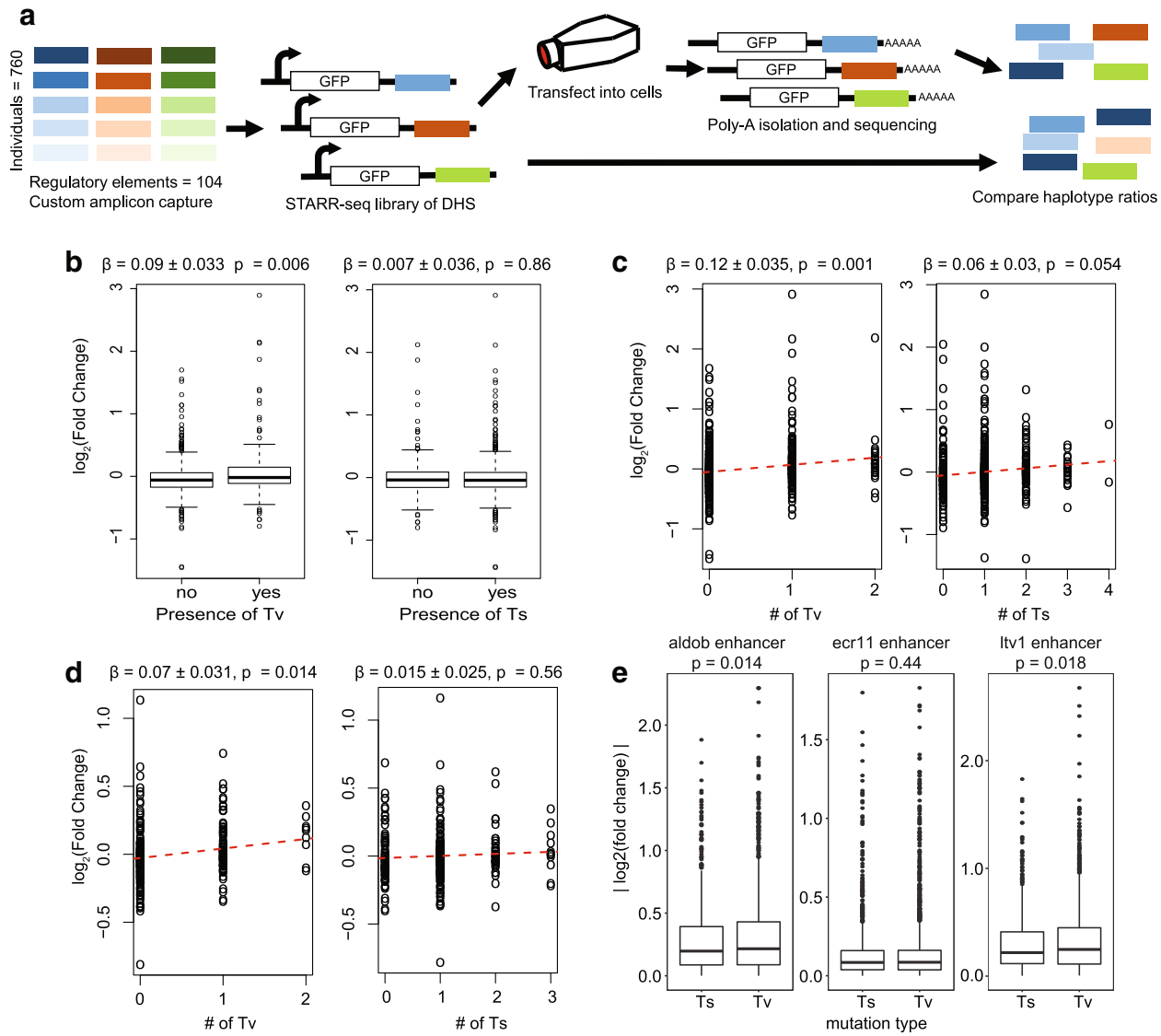

Fig. 4 Tv's have greater regulatory effects than Ts's. **a** Schematic of POP-STARR experimental design. **b** Changes in haplotype effect magnitudes due to the presence of a Ts or Tv within a haplotype. *Black squares* are effects 1.5× of the interquartile range above the upper quartile or below the lower quartile. *P*-values were calculated using a t test. **c** Correlation between change in haplotype effect magnitude and the number of Ts/Tv differences between haplotypes. The *red line* is the regression line. *P*-values were calculated using a t test. **d** Correlation between change in haplotype effect magnitude and the number of Ts/Tv differences between haplotypes which overlap DHS centers. The red line is the regression line. *P*-values were calculated using a t test. **e** *Boxplots* of the effect of transitions and transversions on regulatory element activity as measured in Patwardhan et al., [19]. Each plot shows data from a different enhancer. Each plot shows the absolute $\log_2$(fold change) in regulatory element activity for every Ts and Tv assayed in that enhancer. *P*-values are for a one-sided t test from a linear regression model

was correlated with greater changes in regulatory activity (t test, $\beta = 0.09 \pm 0.033$ s.e.m., $p = 0.006$), while the presence a Ts was not (t test, $\beta = 0.007 \pm 0.036$ s.e.m., $p = 0.86$) (Fig. 4b). This observation suggests that haplotypes within regulatory elements that contain Tv's are more likely to impact activity. Next, we expanded the model to account for the total number of Ts's and Tv's between haplotypes. The total number of Tv's was significantly correlated with the magnitude of changes in regulatory activity (t test, $p = 0.001$) whereas the total number of Ts's was not (t test, $p = 0.054$). Furthermore, the effect of additional Tv's on the magnitude of changes in regulatory element activity was double that of additional Ts's ($\beta = 0.12 \pm 0.035$ s.e.m vs 0.06 ± 0.03 s.e.m.) (Fig. 4c). Since TF binding is enriched at the center of DHS's, we hypothesized that that relative magnitude of effect between Tv's and Ts's would increase near the middle of DHS's. When the same analysis was limited to haplotypes that overlapped the middle third of DHS's, the effect of Tv's was substantially

larger ($\beta = 0.07 \pm 0.031$ s.e.m.), the effect of Ts's was unchanged ($\beta = 0.015 \pm 0.025$ s.e.m.), and the ratio of the effect sizes increased to 4.6-fold (Fig. 4d).

To confirm that our results are not specific to our model system or the 3q25 locus, we performed a similar analysis on a study that used saturation mutagenesis to evaluate the effect of every possible mutation on the activity of three enhancers [19]. In that study, Patwardhan et al. used massively parallel reporter assays to measure the effect of every possible single nucleotide change to three known enhancer regions. The saturation mutagenesis approach allowed for quantification of the effects of every possible mutation on regulatory activity. As in our POP-STARR results, Tv's had greater effects on activity than Ts's, and that difference was greater near the center of each regulatory element (Fig. 4e, Additional file 1: Table S7). Together, these results confirm the results of our POP-STARR assays in an alternative high-throughput reporter system, providing further empirical evidence that Tv's have larger impacts on regulatory element activity than Ts's for both naturally occurring variants and artificially generated mutations.

Discussion

Although the impacts of Ts's and Tv's have been extensively studied in coding sequences, differences in their effects in non-coding DNA has remained largely overlooked. Here, we have shown that there are functional differences in the effects of Ts's and Tv's in non-coding regulatory elements. Specifically, our results show that Tv's are more likely to alter DNA shape, to disrupt TF binding, and to have larger effects on regulatory element activity than Ts's. These findings represent a novel, fundamental property of regulatory variation.

The observed overall effects of Tv's and Ts's on regulatory element activity are modest. That finding is as expected considering earlier results showing that most genetic mutations or variants have modest effects on regulatory element activity [16, 19, 37]. Our estimates of the differences between Ts's and Tv's on regulatory activity are conservative for several reasons. We did not limit our analyses to mutations or variants that influence regulatory activity, nor did we restrict our analysis to binding sites for TFs that are more sensitive to Tv's. We also did not remove from our analysis any regulatory elements with low activity. Our highly inclusive and conservative approach is important to demonstrate that there is a differential overall effect of Tv's and Ts's on regulatory element activity, but leaves to future studies a detailed understanding of the specific circumstances when Tv's disproportionately alter that activity. We expect that elucidating those circumstances will improve prediction of the effects of noncoding variants on both molecular and organismal phenotypes.

Several mechanisms may explain the stronger effects of Tv's. TFs may recognize the purine or pyrimidine structure rather than the specific nucleotide or, alternatively, Tv's may disproportionately alter the DNA backbone, impacting the binding of TFs that recognize backbone shape [39, 40]. One demonstration of that principle comes from the Hox family of TFs that bind DNA by recognizing both sequence and shape independently of each other [25]. Many previous analyses have also suggested that TFs may bind through both sequence direct recognition and indirect recognition [41–45]. Direct recognition occurs when a protein interacts directly with the amino acid sequence of the DNA. Conversely, indirect recognition occurs when proteins interact with the DNA structure. Moreover, some groups have shown that including DNA structure and physiochemical features significantly improves predictions of TF binding across the genome [41, 44]. The importance of DNA shape and structure corroborates the results presented in our study, and provides a potential mechanism explaining why Tv's have larger impacts on TF binding and regulatory activity than Ts's. Understanding those principles of TF recognition may further inform whether specific classes of TFs are particularly impacted by Tv's.

Conclusions

In this work, we demonstrate that transversions have a greater impact on regulatory element activity than transitions. A likely mechanism is that transversions alter the minor groove width and roll of DNA more than transitions, leading to a greater impact on TF binding. These findings provide new insights into the ways that different types of genetic variation can have distinct effects on gene regulation, and suggests that considering whether a variant is a transversion or a transition may be valuable for studying the genetics of gene regulation in many contexts.

Methods

Predicted effects of mutations on DNA shape parameters

To estimate the effects of Ts's and Tv's on DNA shape parameters, we generated a set of random DNA sequences that differed by a single nucleotide, and then used a computational model to predict DNA shape for each sequence. Specifically, we generated 100,000 random 503 bp DNA sequences. We then converted the middle nucleotide in each sequence to all other possible nucleotides, thus yielding 400,000 sequences. We then predicted the minor groove width, roll, propeller twist, and helical twist across each sequence using DNAshapeR [26]. To estimate the effect of Ts's and Tv's on those shape parameters, we summed the absolute difference in each parameter between pairs of sequences that differed by a single Ts or Tv,

respectively. We compared the effect of Ts's and Tv's on each parameter using a linear regression model that included the identity of the starting sequence as a covariate.

Predicted effects of mutations on TF binding

The set of all non-redundant TF binding position frequency matrices (PFMs) were retrieved from the JASPAR database [28]. For each PFM, a pseudocount of 0.1 was added to every element, and the PFM was converted to a position specific scoring matrix (PSSM). The most likely (i.e. consensus) binding sequence was determined for each PSSM. We defined the PSSM score for a given DNA sequence as the sum of the corresponding positions in the PSSM. Then, for each position in each consensus sequence, we calculated the PSSM score of having every possible nucleotide at that position and, subsequently, the change in PSSM score when mutating any nucleotide to any other nucleotide at that position. The final values along with covariates such as the JASPAR motif ID and the position in the motif data were output as a table that was used for statistical analysis. PSSM generation and mutation scoring was performed using BioPython libraries, and statistical analysis was performed in R.

Effects of transitions and transversions in saturation mutagenesis (Patwardhan et al.) dataset

Data from saturation mutagenesis of three regulatory elements were collected [19] and reformatted to give the effect of every possible mutation at every position assayed. Effects from replicate experiments were averaged. A series of linear regression models was then used to evaluate the effect of Ts's and Tv's on regulatory element activity while accounting for differences in regulatory element activity between elements and location of the mutation within each element. The specific models used along with coefficients and test statistics are provided in Additional file 1: Table S7. All analysis was performed using R.

Allele specific binding analysis

We analyzed two publicly available allele-specific binding datasets, one based on the binding of multiple 42 TFs to the diploid personal genome sequence of NA12878, the other based on the binding of CTCF to SNPs discovered through ChIP-seq in 6 different LCLs [30]. We computed the Tv frequency in allele-specific variants and in all variants tested and compared the two frequencies by transforming the assumed binomial distribution to a standard normal and performing a two-tailed Z-test. We pooled allele-specific variants across TFs or across cell lines, making sure to collapse redundant variants. We were ignorant of the overlap of all tested SNPs across cell lines and for simplicity used the mean number of SNPs tested as the null model sample size for that dataset.

Custom amplicon design and capture

Custom amplicon design and capture were performed as described in Vockley, Guo, & Majoros et al. [37] with the following difference: The number of individuals was increased from 95 to 760.

Variant calling and phasing

Variant calling and phasing was performed as described in Vockley, Guo, & Majoros et al. [37] with the following difference: The number of individuals was increased from 95 to 760.

POP-STARR-seq

Population STARR-seq libraries and haplotype effect size calculations were conducted as previously published by Vockley, Guo, & Majoros et al. [37]. In total, we captured 104 DHSs from the genomes of 760 donors via multiplex PCR (Additional files 2, 5 and 6: Tables S1-S3). We sequenced the captured DNA, and called variants using the Genome Analysis Toolkit (GATK) according to Best Practices recommendations [46–48] (Additional file 7: Table S4). The custom amplicon libraries were combined into eight pools (95 individuals per pool) in equimolar ratios. These pools were then amplified and cloned into the STARR-seq backbone. Each pool was transformed into Stellar chemically competent cells per manufacturer protocol. Transformations were recovered for 1 h in SOC medium while shaking (225 rpm at 37 °C) and then incubated for 16 h in 250 mL of LB while shaking (225 rpm at 37 °C). The resulting plasmid reporter input libraries were isolated using a MaxiPrep Kit (Promega). The 8 purified libraries were then pooled in equimolar ratios to create a single plasmid input library. This library was then transfected into T-175 flasks containing HepG2 cells at ~70% confluency with Fugene HD (Promega) at a 5.5:1 ratio of Fugene:DNA. In total, 3 replicate transfections were performed. RNA was harvested after ~48 h Primer sequences for library construction are included in Additional file 1: Table S8.

Comparing effects of Ts's and Tv's on regulatory element activity

To determine the number of Ts's and Tv's between haplotypes, we grouped haplotypes by amplicon. This ensured that each haplotype was compared to only those with the exact same length and start/stop coordinates. For each amplicon, we designated one haplotype at random as the "reference haplotype". For each group of haplotypes, we counted the number of Ts's and Tv's that differed between each haplotype within the group and

the reference haplotype. The change in effect magnitudes between the haplotypes in each amplicon group were calculated as follows: $||\log_2(\textit{effect size of haplotype})| - |\log_2(\textit{effect size of reference haplotype})||$

Amplicon groups which did not contain at least one haplotype with an effect size p-value < 0.05 were excluded from the analysis. Linear regressions were performed using the lm() function in R. When performing regressions we included amplicon number as a variable.

Additional files

Additional file 1: Figure S1. and **Tables S7.** and **S8. Figure S1.** Amplicons targeting DHS and active histone markers in multiple cell lines. In total, 104 DHS were captured using 174 amplicons. Amplicons were tiled across target regions and also captured at least 50 bp upstream and downstream of each DHS. Amplicon are ~400-425 bp in length. **Table S7.** Effects of Tv's on regulatory element activity in Patwardhan et al. dataset. **Table S8.** Population STARR-seq primer sequences. (DOCX 255 kb)

Additional file 2: Table S1. DHS coordinates at 3q25 (BED format). (XLSX 11 kb)

Additional file 3: Table S5. Haplotype Sequences (FASTA format). (XLSX 154 kb)

Additional file 4: Table S6. Haplotype Effects. (XLSX 64 kb)

Additional file 5: Table S2. Custom Amplicon Coordinates (BED format). (XLSX 12 kb)

Additional file 6: Table S3. Custom Amplicon Probe Sequences (BED format). (XLSX 21 kb)

Additional file 7: Table S4. Variant Calls in 760 Individuals (VCF format). (XLSX 2056 kb)

Abbreviations

DHS: DNase-Hypersensitive site; GATK: Genome analysis toolkit; HelT: Helical twist; LCL: Lymphoblastoid cell line; MGW: Minor groove width; POP-STARR: Population-scale STARR-seq; ProT: Propeller twist; PWM: Position weight matrix; SNP: Single nucleotide polymorphism; STARR-seq: Self-transcribing active regulatory region sequencing; TF: Transcription factor; Ts: transition; Tv: transversion; UTR: Untranslated region

Acknowledgments

Not applicable.

Funding

This work was funded by the National Institutes of Health (NIH) DK099820, DK097534, and HD085666. The NIH had no input on the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript.

Availability of data and material

Sequence data reported in this paper are available for download from GEO under accession GSE77743. Data from other aspects of this project is available at the following websites and databases, all of which are publically available and without any specific permissions:
ENCODE Project, https://genome.ucsc.edu/ENCODE/
1000 Genomes Database, http://browser.1000genomes.org/index.html
Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/
UCSC Genome Browser, http://genome.ucsc.edu
JASPAR, http://jaspar.genereg.net/
Analysis for JASPAR dataset, https://github.com/ReddyLab/TransversionsInRegElements
Analysis of Patwardhan et al., https://github.com/ReddyLab/TransversionsInRegElements
VCFtools, https://vcftools.github.io/man_latest.html
BEDtools, http://bedtools.readthedocs.org/en/latest/

Authors' contributions

CG and TER conceived the study. CG performed the POP-STARR assay and all related analysis. ICM performed the allele-specific binding analysis. TER analyzes Ts:Tv in the JASPAR database. DMS and MN called the variants in the initial sequencing. ASA assisted with all analyses. CG, TER, and WLL designed the sequencing study. All authors contributed to the writing and editing of the final manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Publisher's Note

Author details

[1]Center for Genomic and Computational Biology, Duke University Medical School, Durham, NC 27710, USA. [2]University Program in Genetics and Genomics, Duke University, Durham, NC 27710, USA. [3]Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27710, USA. [4]Department of Preventive Medicine, Division of Biostatistics, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA. [5]Center for Statistical Genetics and Genomics, Duke University Durham, North Carolina 27710, USA. [6]Department of Biostatistics and Bioinformatics, Duke University Medical School, Durham, NC 27710, USA. [7]Division of Endocrinology, Metabolism and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA. [8]Present Address: Biostatistics & Bioinformatics, 101 Science Dr., 2347 CIEMAS, Durham, NC 27708, USA.

Received: 26 August 2016 Accepted: 10 May 2017
Published online: 19 May 2017

References

1. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
2. Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE, Myers RM. Distinct properties of cell-type-specific and shared transcription factor binding sites. Mol Cell. 2013;52:25–36.
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
4. Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, Myers RM. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome Res. 2009;19:2163–71.
5. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014;95:535–52.
6. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518:337–43.
7. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
8. Vockley CM, Barrera A, Reddy TE. Decoding the role of regulatory element polymorphisms in complex disease. Curr Opin Genet Dev. 2016;43:38–45.
9. Lowe WL, Reddy TE. Genomic approaches for understanding the genetics of complex disease. Genome Res. 2015;25:1432–41.
10. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60.
11. McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010;328:235–9.

12. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet. 2010;11:533–8.
13. Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res. 2012;22:860–9.
14. Soccio RE, Chen ER, Rajapurkar SR, Safabakhsh P, Marinis JM, Dispirito JR, Emmett MJ, Briggs ER, Fang B, Everett LJ, et al. Genetic variation determines PPARgamma function and anti-diabetic drug response in vivo. Cell. 2015; 162:33–44.
15. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010;466:714–9.
16. Guo C, Ludvik AE, Arlotto ME, Hayes MG, Armstrong LL, Scholtens DM, Brown CD, Newgard CB, Becker TC, Layden BT, et al. Coordinated regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase HKDC1. Nat Commun. 2015;6:6069.
17. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci U S A. 2012;109:19498–503.
18. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan Jr CG, Kinney JB, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012;30:271–7.
19. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012;30:265–70.
20. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS, Sankaran VG. Systematic functional dissection of common genetic variation affecting red blood cell traits cell. 2016;165(6):1530–45.
21. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 1999;22:231–8.
22. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20.
23. Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 1985;2:150–74.
24. Roman TS, Marvelle AF, Fogarty MP, Vadlamudi S, Gonzalez AJ, Buchkovich ML, Huyghe JR, Fuchsberger C, Jackson AU, Wu Y, et al. Multiple hepatic regulatory variants at the GALNT2 GWAS locus associated with high-density lipoprotein cholesterol. Am J Hum Genet. 2015;97:801–15.
25. Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, Rohs R, Mann RS. Deconvolving the recognition of DNA shape from sequence. Cell. 2015;161:307–18.
26. Chiu TP, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNAshapeR: an R/ Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics. 2016;32:1211–3.
27. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordan R, Rohs R. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci U S A. 2015;112:4654–9.
28. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2015;44:110–5.
29. Stormo GD, Fields DS. Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci. 1998;23:109–13.
30. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun. 2016;7:11101.
31. Freathy RM, Mook-Kanamori DO, Sovio U, Prokopenko I, Timpson NJ, Berry DJ, Warrington NM, Widen E, Hottenga JJ, Kaakinen M, et al. Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. Nat Genet. 2010;42:430–5.
32. Horikoshi M, Yaghootkar H, Mook-Kanamori DO, Sovio U, Taal HR, Hennig BJ, Bradfield JP, St Pourcain B, Evans DM, Charoen P, et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. Nat Genet. 2013;45:76–82.
33. Urbanek M, Hayes MG, Armstrong LL, Morrison J, Lowe LP, Badon SE, Scheftner D, Pluzhnikov A, Levine D, Laurie CC, et al. The chromosome 3q25 genomic region is associated with measures of adiposity in newborns in a multi-ethnic genome-wide association study. Hum Mol Genet. 2013;22: 3583–96.
34. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22:1748–59.
35. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal lari R, Lupien M, Markowitz S, Scacheri PC. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 2014;24:1–13.
36. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol. 2012;30:1095–106.
37. Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, Lowe Jr WL, Reddy TE. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res. 2015;25:1206–14.
38. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013; 339:1074–7.
39. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein-DNA recognition. Nature. 2009;461:1248–53.
40. Mathelier A, Xin B, Chiu TP, Yang L, Rohs R, Wasserman WW. DNA shape features improve transcription factor binding site predictions in vivo. Cell Syst. 2016;3(278–286):e274.
41. Bauer AL, Hlavacek WS, Unkefer PJ, Mu F. Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. PLoS Comput Biol. 2010;6:e1001007.
42. Dai Z, Guo D, Dai X, Xiong Y. Genome-wide analysis of transcription factor binding sites and their characteristic DNA structures. BMC Genomics. 2015; 16 Suppl 3:S8.
43. Tsai ZT, Shiu SH, Tsai HK. Contribution of sequence motif, chromatin State, and DNA structure features to predictive models of transcription factor binding in yeast. PLoS Comput Biol. 2015;11:e1004418.
44. Maienschein-Cline M, Dinner AR, Hlavacek WS, Mu F. Improved predictions of transcription factor binding sites using physicochemical features of DNA. Nucleic Acids Res. 2012;40:e175.
45. Steffen NR, Murphy SD, Tolleri L, Hatfield GW, Lathrop RH. DNA sequence and structure: direct and indirect recognition in protein-DNA binding. Bioinformatics. 2002;18 Suppl 1:S22–30.
46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
47. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
48. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics/editoral board, Andreas D. Baxevanis… [et al.]. 2013;11, 11 10 11–11 10 33.