Published in final edited form as:

Hum Genet. 2016 February ; 135(2): 223–232. doi:10.1007/s00439-015-1628-4.

Linking Short Tandem Repeat Polymorphisms with Cytosine Modifications in Human Lymphoblastoid Cell Lines

Zhou Zhang^{1,2}, Yinan Zheng^{2,3}, Xu Zhang⁴, Cong Liu⁵, Brian Thomas Joyce^{2,6}, Warren A. Kibbe⁷, Lifang Hou^{2,8,†}, and Wei Zhang^{2,8,9,*;†}

¹Driskill Graduate Program in Life Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL 60611

²Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611

³Institute for Public Health and Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611

⁴Section of Hematology/Oncology, Department of Medicine, University of Illinois at Chicago, Chicago, IL 60612

⁵Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60612

⁶Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, IL 60612

⁷Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD 20850

⁸The Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL 60611

⁹Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611

Abstract

Inter-individual variation in cytosine modifications has been linked to complex traits in humans. Cytosine modification variation is partially controlled by single nucleotide polymorphisms (SNPs), known as modified cytosine quantitative trait loci (mQTL). However, little is known about the role of short tandem repeat polymorphisms (STRPs), a class of structural genetic variants, in regulating cytosine modifications. Utilizing the published data on the International HapMap Project lymphoblastoid cell lines (LCLs), we assessed the relationships between 721 STRPs and the modification levels of 283540 autosomal CpG sites. Our findings suggest that, in contrast to the predominant *cis*-acting mode for SNP-based mQTL, STRPs are associated with cytosine modification levels in both *cis*-acting (local) and *trans*-acting (distant) modes. In local scans within the ± 1 Mb windows of target CpGs, 21, 9, and 21 *cis*-acting STRP-based mQTL

*Corresponding author: 680 N. Lake Shore Dr., Suite 1400. Chicago, IL 60611; Tel: +1 312-503-1040; Fax: +1 312-908-9588; wei.zhang1@northwestern.edu.

[†]These authors contributed equally to this work as senior authors.

were detected in CEU (Caucasian residents from Utah, USA), YRI (Yoruba people from Ibadan, Nigeria), and combined samples, respectively. In contrast, 139420, 76817, and 121866 *trans*-acting STRP-based mQTL were identified in CEU, YRI, and combined samples, respectively. A substantial proportion of CpG sites detected with local STRP-based mQTL were not associated with SNP-based mQTL, suggesting that STRPs represent an independent class of mQTL. Functionally, genetic variants neighboring CpG-associated STRPs are enriched with genome-wide association study (GWAS) loci for a variety of complex traits and diseases, including cancers, based on the National Human Genome Research Institute (NHGRI) GWAS Catalog. Therefore, elucidating these STRP-based mQTL in addition to SNP-based mQTL can provide novel insights into the genetic architectures of complex traits.

Keywords

short tandem repeat polymorphism; cytosine modification; DNA methylation; lymphoblastoid cell line; quantitative trait locus; complex trait

INTRODUCTION

Genome-wide association studies (GWAS) have implicated genetic variants, particularly SNPs (single nucleotide polymorphisms) in human complex traits, including risks for common diseases, drug response, and quantitative gene expression phenotypes (Hindorff et al. 2009). For example, previous studies have identified common genetic variants associated with gene expression, known as eQTL (expression quantitative trait loci) (Duan et al. 2008; Morley et al. 2004; Spielman et al. 2007; Stranger et al. 2007; Zhang et al. 2008). However, GWAS typically rely on genotyping of SNPs, the most common form of genetic variants in the human genome. Besides SNPs, increasing attention has been targeting structural variants such as copy number variants (CNVs) (Conrad et al. 2010; Cooper et al. 2011; McCarroll et al. 2008; Perry 2008), yet little is known about the relationships between structural variants and complex traits.

Over half of the human genome is comprised of various repetitive and repeat-derived DNA elements (The International Human Genome Sequencing Consortium 2001). In particular, short tandem repeats (STRs), also known as microsatellites or simple tandem repeats, which repeat multiple times in a head-to-tail manner of 2–5 base pairs, have been reported to account for 3% of the human genome (Ellegren 2004). The length of STRs can be highly polymorphic among different individuals and populations (Payseur and Jing 2009; Payseur et al. 2011; Payseur et al. 2008). In addition, STRs feature high mutation rates as well (Ellegren 2000), with a mutation rate of 10^{-3} to 10^{-5} per generation (Pumpernik et al. 2008; Weber and Wong 1993). Although serving as an important source of genetic variation (Mills et al. 2006), STRs remain poorly studied due to technical difficulties of capturing highly variable and non-unique parts of the genome (Brahmachary et al. 2014).

Besides genetic variation, epigenetic systems, such as the methylation of 5'-cytosine at CpG dinucleotides in DNA, also contribute to gene expression (Murrell et al. 2004; Pai et al. 2011; Zhang et al. 2014). Cytosine modification is known to silence gene expression in *cis*-regulation (Siegfried et al. 1999). Recent mQTL (modified cytosine quantitative trait loci)

mapping studies suggest that cytosine modifications are partially controlled by genetic variation (Gamazon et al. 2013; Zhang et al. 2010). SNPs associated with both cytosine modification and gene expression levels were identified in previous mQTL and eQTL mapping studies (Bell et al. 2011; Zhang et al. 2014), suggesting that SNPs may underlie the links between epigenetic traits and gene expression phenotypes. However, as an important source of natural genetic variations, STRs' relationships with cytosine modification levels remain to be characterized. Understanding their relationships will provide us a more comprehensive picture of the genetic basis for cytosine modifications, thus enhancing our knowledge of human complex traits.

Human lymphoblastoid cell lines (LCLs), a widely used human genetics model (Payseur et al. 2008; Zhang et al. 2014), have been investigated by our group and other investigators for eQTL, SNP-based mQTL, and more recently pQTL (protein quantitative trait loci) (Duan et al. 2008; Moen et al. 2013; Stark et al. 2014; Zhang et al. 2008; Zhang et al. 2015; Zhang et al. 2014). In the current study, we integrated the cytosine modification levels of 283540 autosomal CpG sites and the repeat length data of 721 autosomal STRs in 117 unrelated LCLs samples from the International HapMap Project (Payseur et al. 2008; The International HapMap Consortium 2003, 2005, 2007, 2010). We assessed the associations between STRs and cytosine modifications to characterize the STRP-based mQTL, and gained novel biological insights into their potential implications on human complex traits.

MATERIALS AND METHODS

Figure 1 shows the general workflow of this study. Briefly, STRP-based mQTL were identified by assessing the relationships between STRPs and cytosine modifications in two HapMap populations. Then, STRP-based mQTL were compared with SNP-based mQTL. Enrichment analysis of KEGG pathways and GWAS loci were performed for cis- and trans-acting STRP-based mQTL, respectively.

Obtaining cytosine modification data

The cytosine modification data, profiled using the Illumina HumanMethylation450 BeadChip (450K array), on 133 unrelated HapMap LCL samples (60 CEU samples: Caucasian residents from Utah, USA and 73 YRI samples: Yoruba people from Ibadan, Nigeria) were previously generated by our group (Moen et al. 2013). The raw data can be downloaded from the NCBI Gene Expression Omnibus (GEO Accession Number: GSE39672). Data processing was described in our previous publication (Moen et al. 2013). Briefly, we removed those CpG probes that: (i) ambiguously mapped to the human genome (Zhang et al. 2012); (ii) contained common SNPs if the SNPs located within 20 bp from interrogated CpG sites had minor allele frequency (MAF) > 0.01 (based on dbSNP v135) (Sherry et al. 2001); (iii) had a low calling rate (defined as 6 out of 133 samples with a detection p -value < 0.01); or (iv) were on sex chromosomes. The final dataset was comprised of 283540 highly reliable, autosomal CpG sites (Zhang et al. 2014; Zhang et al. 2012). Notably, given the short range of potential co-modification effect (Moen et al. 2013) and the design of the 450K array (~20 CpGs per known gene) (Bibikova et al. 2011), it is expected that the analytical set of 283540 CpGs do not contain a substantial number of co-modified

CpG sites. The *M*-value, defined as the $\log_2$ ratio of the intensities of methylated probes versus unmethylated probes (Du et al. 2010), was summarized for each CpG site in each individual. The *M*-values were quantile normalized and COMBAT corrected for batch effect across all samples (Bolstad et al. 2003; Johnson et al. 2007).

Obtaining STRP data

We used published STRP data from 117 unrelated HapMap LCL samples (59 CEU and 58 YRI) genotyped from Marshfield 5 cM genomic linkage screening sets (Payseur et al. 2008). This dataset was comprised of 721 autosomal STRPs, which were mapped to the human genome reference at the UCSC website (hg17, Build 35). Of these 721 STRPs, 51 were di-nucleotide repeats, 149 were tri-nucleotide repeats, 511 were tetra-nucleotide repeats, and 10 were penta-nucleotide repeats. In order to be compatible with the cytosine modification data, we performed coordinate transformation for the STRP data, and lift it over from hg17 (build 35) to hg19 (build 37) human genome coordinates. In total, 59 CEU and 58 YRI samples, on which we have both STRP data and cytosine modification levels were included for further analysis.

Identification of STRP-based mQTL

To detect local mQTL, correlations between the length of STRPs and the cytosine modification levels (within the ± 1 Mb windows of the STRPs) were evaluated. A linear model was fitted for CEU and YRI populations separately: cytosine modification level $\sim$ length of STRPs + gender + error. This model was used to assess the association between cytosine modification levels and STRP length, and to detect different association patterns between the two populations. Total of 137984 CpGs and 721 STRPs (159779 STRP-CpG pairs) were tested in both populations. Adjusted *p*-value was estimated by 1000 permutations for local scans following the max *T* procedure (Westfall and Young 1993). An adjusted *p*-value < 0.05 (corresponding to nominal *p*-value $< 4.0 \times 10^{-5}$) was considered significant.

For whole-genome scan, the same model was fitted, and a nominal *p*-value $< 10^{-5}$ was considered significant. 283540 CpGs and 721 STRPs were separately tested in CEU, YRI, and combined samples. Further, for those STRPs that correlated with over 200 CpGs (defined as master STRPs), we identified the genes within the windows of ± 10 Kb from these STRPs. For these genes, KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto 2000) pathway enrichment analysis was performed in each population using hypergeometric test. Gene sets containing a minimum of three genes with a Benjamini-Hochberg (Benjamini and Hochberg 1995) adjusted *p*-value < 0.05 were reported.

Evaluating gene regulatory functions of STPR-based mQTL

For the significant STRP-associated local CpG sites, we evaluated the within-population correlation between gene expression levels and cytosine modification levels at these sites. The Affymetrix Human Exon Array 1.0ST was previously used to profile gene expression in the same CEU and YRI samples (GEO Accession Number: GSE9703) (Zhang et al. 2008). We followed the procedure described in our previous publication (Zhang et al. 2014).

Briefly, probe sequences were aligned to the human genome (build 37), allowing 1 mismatch. Probes with perfect, unique alignments were further filtered by excluding probes containing common SNPs (MAF > 0.05) based on dbSNP (Sherry et al. 2001) v135 and probes interrogating multiple genes. Probe intensities were $\log_2$ transformed, background corrected (Borevitz et al. 2003) and quantile normalized. Probe intensity was subtracted by the corresponding probe mean across samples. Gene-level expression intensities were summarized as mean probe intensity within each gene. Genes within ± 10 Kb windows of significant STRP-associated local CpGs were included in the analysis. In total, 21 and 12 autosomal genes with unique Entrez Gene IDs were included in the analyses for CEU and YRI, respectively. Gene expression levels were then regressed on cytosine modification levels. An adjusted p -value < 0.05 following the Benjamini-Hochberg procedure was considered significant.

Comparison between STRP-based mQTL and SNP-based mQTL

To examine cytosine modification variation that cannot be fully explained by SNPs, CpGs that showed significant associations with STRPs in two populations in local scans were compared with our previously detected SNP-based mQTL maintained at the SNP and Copy Number Annotation (SCAN) database (<http://www.scandb.org>) (Zhang et al. 2015). First we identified SNPs that have been genotyped within the windows of ± 1 Mb from the significant CpGs, i.e., those CpG sites showing associations with STRPs in CEU and YRI, separately. Then, we used the SCAN database to identify which STRP-associated local CpGs' modification levels are predicted by these SNPs with a nominal p -value less than 10^{-4} . These CpG sites were then compared with our local STRP-based mQTL-associated CpGs.

Enrichment of GWAS loci among STRP-based mQTL

The NHGRI GWAS Catalog has collected published GWAS results for > 600 complex traits, including risks for common diseases (Hindorff et al. 2009). We obtained the NHGRI GWAS loci (released on April 15, 2015) reported to be associated with complex traits with a p -value < 1×10^{-8} . SNP genotypic data were downloaded from the International HapMap Project (The International HapMap Consortium 2007) (release 27). Those SNPs that deviated from the Hardy-Weinberg equilibrium (p -value < 10^{-4}) within a population were removed. The analytic set was comprised of > 2 million common SNPs with MAF > 0.05 that were genotyped in at least 48 samples in each population. As the majority of published GWAS were carried out in Caucasian samples, we chose to use the CEU samples in this analysis. To assess the enrichment of GWAS loci among local STRP-based mQTL (at both adjusted p -value < 0.05, and nominal p -value < 10^{-3}) from different distances (e.g., 100Kb, 500Kb, and 1Mb), the null distributions were generated by overlapping GWAS loci with randomly sampled HapMap SNPs 10000 times (the number of SNPs for each sampling was determined by the number of SNPs within different windows of *cis*-acting STRP-based mQTL). The observed true numbers of GWAS loci within different windows were compared with the null distributions.

Co-localization of *cis*-acting STRP-based mQTL with ENCODE enhancers

To explore the potential mechanisms of cytosine modification regulation by STRPs, we characterized the co-localization of identified *cis*-acting STRPs in CEU population with the transcription factor binding sites (TFBS) and enhancers from the Encyclopedia of DNA Elements (ENCODE) Project (Encode Project Consortium 2012), on a CEU cell line GM12878 (Ernst et al. 2011). Our identified *cis*-acting STRPs were mapped to the windows of ± 10 Kb from the center of each TFBS or enhancer. Human genome reference (hg19) were used for both ENCODE and STRPs data.

RESULTS

Influence of STRPs on local cytosine modifications

Pearson's correlations (ρ) of STRP length and M -values of local CpGs within ± 1 Mb windows of STRPs for CEU, YRI, and the combined samples were calculated. Most analyzed STRP-CpG pairs showed relatively low correlation ($\rho < 0.4$) within ± 1 Mb windows (Supplemental Figure 1). By fitting linear models of cytosine modification levels on STRP length, we identified 183, 173, and 176 STRP-CpG association relationships in CEU, YRI, and the combined samples, respectively (p -value $< 10^{-3}$; Table 1, Supplemental Table 1). In contrast, at an adjusted p -value < 0.05 (corresponding to nominal p -value $< 4.0 \times 10^{-5}$), in total 21, 9, and 21 STRP-CpG association relationships for CEU, YRI, and the combined samples were identified, respectively (Table 1). Among the 21 STRP-base mQTL (adjusted p -value < 0.05) in CEU, 15 of them were co-localized with TFBS/enhancers from ENCODE (Supplemental Table 1).

Influence of STRPs on distant cytosine modifications

To evaluate the effects of STRPs on levels of distant cytosine modifications, we also fitted the linear regression models between the length data of 721 STRs and the M -values of 283540 CpGs across the whole genome. The observed p -value significantly differed from theoretical p -values from the cutoff of 10^{-5} (Supplemental Figure 2), we therefore used the threshold of nominal p -value = 10^{-5} , for CEU, YRI, and combined samples in whole-genome scan. We identified overall 74676 CpGs in CEU, YRI and combined samples (p -value $< 10^{-5}$; Table 1, Supplemental Table 2, Supplemental Figure 2). The numbers of significant STRP-CpG associations we observed in CEU, YRI, and combined samples were 139420, 76817, and 121866, respectively. Out of these associations, there were 41794, 30994, and 50104 unique CpGs in CEU, YRI, and combined samples, respectively.

We identified 177 and 103 master STRPs (i.e., those with > 200 associated CpG sites) in CEU, YRI, respectively. For genes within ± 10 Kb windows of these master STRPs, KEGG pathway enrichment analysis were performed. Gene sets containing a minimum of three genes with a Benjamini-Hochberg adjusted p -value < 0.05 for KEGG were reported (Supplemental Table 3).

Comparison of STRP-based mQTL between CEU and YRI

For the local scan, with a nominal p -value cutoff of 10^{-3} , there were 183 and 173 significant STRP-CpG association pairs in CEU and YRI, respectively. However, there were only two

common pairs among the two populations (Fig. 2). At a more stringent cutoff of adjusted p -value < 0.05 (corresponding to nominal p -value $< 4.0 \times 10^{-5}$), the significant STRP-CpG association pairs were 21 and 9 in CEU and YRI, respectively, with one STRP-CpG pair observed in both populations (Fig. 2). For the whole genome scan, a total of 17042 STRP-CpG common associations for 9858 unique CpG sites were detected in CEU and YRI. The chromosomal distribution of STRP-based mQTL showed different pattern between the two populations (Fig. 3).

Comparison of STRP-based and SNP-based mQTL

SNP-based mQTL within ± 1 Mb windows of CpGs, which were significantly associated with local STRPs, were identified from the SCAN database (Zhang et al. 2015). Within the ± 1 Mb windows of target CpGs, 40 out of 183 CpGs (21.9%) in CEU, and 39 out of 173 CpGs (22.0%) in YRI were overlapped with previously detected SNP-based mQTL (p -value $< 10^{-3}$; Fig. 2, Supplemental Table 1). In contrast, 13 out of 21 CpGs in CEU, and 4 out of 9 CpGs in YRI were overlapped with SNP-based mQTL (adjusted p -value < 0.05 , corresponding to nominal p -value $< 4.0 \times 10^{-5}$; Fig. 2, Supplemental Table 1).

Enrichment of GWAS loci among *cis*-acting STRP-based mQTL

To assess whether the GWAS loci are enriched with *cis*-acting STRP-based mQTL, we obtained 4807 unique SNPs, covering 685 unique diseases/traits from the NHGRI GWAS Catalog, which had a nominal association p -value $< 1 \times 10^{-8}$ (Hindorff et al. 2009). The GWAS loci < 1 Mb away from local STRP-based mQTL were analyzed. We identified 120 unique GWAS loci, covering 75 (10.9%) unique diseases/traits. These SNPs were within ± 1 Mb windows of 18 *cis*-acting STRP-based mQTL (adjusted p -value < 0.05 ; Supplemental Table 4). Overall, 2.5% of the unique GWAS loci were annotated, covering 10.9% of the investigated diseases/traits. In comparison with the null distribution in ± 500 Kb-1 Mb windows, the GWAS loci were enriched among *cis*-acting STRP-based mQTL at different p -value thresholds (adjusted p -value < 0.05 , Fig. 4; nominal p -value $< 10^{-3}$, Supplemental Figure 3).

DISCUSSION

We assessed the association relationships of 721 STRPs and cytosine modifications for 283540 autosomal CpG sites in two HapMap populations. Our observations suggest that, in contrast to the predominant *cis*-acting mode for SNP-based mQTL (Zhang et al. 2014), STRPs are associated with cytosine modification levels in both *cis*-acting (local) and *trans*-acting (distant) modes (Fig. 3). The correlation of STRP-CpG effect size and the distance between significant STRP-CpG was low ($r=0.065$), suggesting STRPs' contribution to cytosine modification variation is not sensitive to distance, a dramatic difference from previously reported SNP-based mQTL (Zhang et al. 2014). Both *cis*- and *trans*-acting STRP-based mQTL showed a population-specific pattern (Figs. 2, 3), which is consistent with our previous findings in SNP-based mQTL, suggesting that STRPs are important contributors to the population specificity of mQTL effects (Fraser et al. 2012; Moen et al. 2013; Zhang et al. 2014). However, given that a substantial proportion of STRP-based

mQTL were not overlapped with SNP-based mQTL, these two types of mQTL could contribute independently to population-specificity.

In comparison with the SNP-based mQTL detected from the same samples, a large proportion of the CpG sites detected with STRP-based mQTL appeared to be associated with STRPs alone. Therefore, STRPs may represent an independent source of genetic basis for cytosine modifications. Although a substantial proportion of the CpG sites were solely associated with STRPs, some of the detected CpG sites could be affected by both STRP and SNPs (Fig. 2, Supplemental Table 1). For example, cg08922148, the common STRP-associated local CpG site observed in both CEU and YRI, was associated with 37 SNPs according to our previous study (Zhang et al. 2014). This finding suggests more complicated mechanisms may underlie the regulation of cytosine modification, and STRPs may be a complementary to SNP-based mQTL. Adding this class of genetic variations will enhance our knowledge of the genetic basis for cytosine modifications.

We evaluated the associations between gene expression levels (within ± 10 Kb windows of STRP-associated local CpGs) and modification levels of STRP-associated local CpGs. However, no significant result was found after FDR adjustment ($FDR < 0.05$; Supplemental Table 5). This finding is consistent with previous study that SNP-based mQTL and eQTL may provide orthogonal ways of functionally annotating genetic variation (Gamazon et al. 2013). However, at 20% FDR level, one CpG site, cg13171904, showed association with *TLN2* in the CEU population. *TLN2* is a protein coding gene related to actin binding and structural constituent of cytoskeleton (Monkley et al. 2001). Both cg13171904 and *TLN2* are located on chromosome 15. This gene has been linked to Alzheimer's disease according to the NHGRI GWAS Catalog (Hindorff et al. 2009; Stein et al. 2010). Interestingly, Alzheimer's disease has been reported to relate to STRPs located on chromosome 14 (St George-Hyslop et al. 1992). Although we did not observe any STRPs from chromosome 14 associated with cg13171904, this may due to the small number of STRPs we investigated. This still might represent an example where STRPs regulate cytosine modifications in a distant mode, which further affect gene expression levels and lead to diseases/traits. Besides this potential link from STRPs to neural diseases through cytosine modifications, several nervous system-related pathways were enriched in KEGG pathway analysis for genes close to master STRPs (i.e., those with > 200 associated CpG sites) in CEU and YRI (Supplemental Table 3). For example, long-term depression (LTD), an activity-dependent reduction in the efficacy of neuronal synapses lasting hours or longer (Kuroda et al. 2001), was enriched in both CEU and YRI. This finding is consistent with a previous study that LTD is linked to STRP (Bolton et al. 2013), suggesting these master STRPs may have functional implications for complex diseases/traits.

GWAS to date have generated a large number of disease/trait associations for which biological interpretation is still lacking. Our previous study has demonstrated that GWAS loci were enriched for SNP-based mQTL (Zhang et al. 2014), suggesting that linking trait-associated loci with intermediate phenotypes such as cytosine modification likely provided further insights into GWAS results (Zhang et al. 2011). Here, we found that GWAS loci were enriched within ± 500 Kb – 1Mb windows of *cis*-acting STRP-based mQTL, suggesting that GWAS loci are enriched in a relatively far distance from *cis*-acting STRP-

based mQTL. This may be caused by long-range linkage disequilibrium between STRPs and SNPs across the human genome (Payseur et al. 2008). The identified GWAS loci (within $\pm$ 1Mb windows of *cis*-acting STRP-based mQTL) were annotated with a broad range of disease categories, such as breast cancer (Wooster et al. 1994), colorectal cancer, and schizophrenia (Berto et al. 2007; Hattori et al. 2001). These diseases have already been demonstrated to relate to STRPs (Berto et al. 2007; Hattori et al. 2001; Wooster et al. 1994). For example, GATA4H01-cg08922148 association we identified was related *CIT* gene. *CIT* gene, which encodes citron rho-interacting serine/threonine kinase, is involved in central nervous system development, and is associated with schizophrenia, and bipolar disorder (Berto et al. 2007; Hattori et al. 2001). Other annotated diseases, including Crohn's disease, inflammatory bowel disease, and chronic lymphocytic leukemia (Supplemental Table 4), are related to the functions of B lymphocytes from which LCLs are derived. This observation suggests that the cell lineage specificity of cytosine modifications plays a critical role in determining disease outcomes. Overall, these identified GWAS loci represent about 2.5% of unique SNPs from the NHGRI GWAS Catalog. This relatively small number is likely due to the small number of STRs genotyped in the currently available dataset. In contrast, these loci correspond to 10.9% of the investigated complex traits in the NHGRI GWAS Catalog.

Although the mechanism underlying *trans*-acting effects of STRPs on cytosine modification is not clear, one possible explanation for the *cis*-acting STRP-based mQTL could be their co-localization with TFBS/enhancers (Supplemental Table 1). As a recent report has demonstrated, transcription factors (TFs) can induce hypomethylation, even when these sites are methylated; on the other hand, departure of TFs from decommissioned enhancers allows for subsequent re-methylation (Stadler et al. 2011). Thus, with a mutation rate of 10^{-3} to 10^{-5} per generation (Pumpernik et al. 2008; Weber and Wong 1993), STRPs could affect the accessibility of co-localized TFBS/enhancers, further alter local cytosine modifications. However, future work is needed to uncover the molecular mechanisms under which STRPs regulate DNA methylation patterns through changing the accessibility of TFBS/enhancers.

We recognize that our study is subject to several limitations. First, because of the relatively small sample size, we were not able to fully exclude false negatives or false positive findings. Second, the STRP markers available to us were not comprehensive enough to capture the massive genome-wide STR variations, in comparison with over 2 million common SNPs in previous SNP-based mQTL mapping studies (The International HapMap Consortium 2007, 2010; Zhang et al. 2014). However, these limitations are inherent in STRPs studies due to highly variable and non-unique nature of the STRs (Brahmachary et al. 2014). In future research, this obstacle may be overcome with the development of long-read next-generation sequencing technologies (Li et al. 2015; Ram et al. 2015).

In conclusion, STRPs appear to be an independent class of genetic variants that regulate the levels of cytosine modifications in both *cis*- and *trans*-manner in a population-specific way. Our findings provide a new angle to enhance our understanding of the genetic architectures of epigenetic variations. Elucidating STRP-based mQTL may provide novel insights into the underlying biological mechanisms of human complex traits and disease pathogenesis. Inclusion of STRPs in future GWAS is highly recommended for a comprehensive capture of genetic variation. We expect that mapping STRP-based mQTL using diverse cell types and

treatment conditions will greatly expand our understanding of human complex traits in general.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was partially supported by grants from the National Institutes of Health: R21HG006367 (to WZ), R21CA187869 (to WZ and LH), and The Robert H. Lurie Comprehensive Cancer Center-Developmental Funds P30CA060553 (to WZ). Authors would like to thank Dr. Lei Liu for helpful discussions.

REFERENCES

- Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* 2011; 12:R10.10.1186/gb-2011-12-1-r10 [PubMed: 21251332]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple Testing. *J Roy Stat Soc B Met.* 1995; 57:289–300.10.2307/2346101
- Berto G, Camera P, Fusco C, Imarisio S, Ambrogio C, Chiarle R, et al. The Down syndrome critical region protein TTC3 inhibits neuronal differentiation via RhoA and Citron kinase. *J Cell Sci.* 2007; 120:1859–1867.10.1242/jcs.000703 [PubMed: 17488780]
- Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. *Genomics.* 2011; 98:288–295.10.1016/j.ygeno.2011.07.007 [PubMed: 21839163]
- Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003; 19:185–193. [PubMed: 12538238]
- Bolton KA, Ross JP, Grice DM, Bowden NA, Holliday EG, Avery-Kiejda KA, et al. STaRRRT: a table of short tandem repeats in regulatory regions of the human genome. *BMC Genomics.* 2013; 14:795.10.1186/1471-2164-14-795 [PubMed: 24228761]
- Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, et al. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* 2003; 13:513–523.10.1101/gr.541303 [PubMed: 12618383]
- Brahmachary M, Guilmatre A, Quilez J, Hasson D, Borel C, Warburton P, et al. Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. *PLoS Genet.* 2014; 10:e1004418.10.1371/journal.pgen.1004418 [PubMed: 24945355]
- Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. *Nature.* 2010; 464:704–712.10.1038/nature08516 [PubMed: 19812545]
- Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. *Nat Genet.* 2011; 43:838–846.10.1038/ng.909 [PubMed: 21841781]
- Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics.* 2010; 11:587.10.1186/1471-2105-11-587 [PubMed: 21118553]
- Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, Clark TA, et al. Genetic architecture of transcript-level variation in humans. *Am J Hum Genet.* 2008; 82:1101–1113.10.1016/j.ajhg.2008.03.006 [PubMed: 18439551]
- Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet.* 2000; 24:400–402.10.1038/74249 [PubMed: 10742106]

- Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 2004; 5:435–445.10.1038/nrg1348 [PubMed: 15153996]
- Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74.10.1038/nature11247 [PubMed: 22955616]
- Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.* 2011; 473:43–49.10.1038/nature09906 [PubMed: 21441907]
- Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. *Genome Biol.* 2012; 13:R8.10.1186/gb-2012-13-2-r8 [PubMed: 22322129]
- Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, Cox NJ, et al. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry.* 2013; 18:340–346.10.1038/mp.2011.174 [PubMed: 22212596]
- Hattori E, Ebihara M, Yamada K, Ohba H, Shibuya H, Yoshikawa T. Identification of a compound short tandem repeat stretch in the 5'-upstream region of the cholecystokinin gene, and its association with panic disorder but not with schizophrenia. *Mol Psychiatry.* 2001; 6:465–470.10.1038/sj.mp.4000875 [PubMed: 11443535]
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 2009; 106:9362–9367.10.1073/pnas.0903103106 [PubMed: 19474294]
- Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007; 8:118–127.10.1093/biostatistics/kxj037 [PubMed: 16632515]
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000; 28:27–30. [PubMed: 10592173]
- Kuroda S, Schweighofer N, Kawato M. Exploration of signal transduction pathways in cerebellar long-term depression by kinetic simulation. *J Neurosci.* 2001; 21:5693–5702. [PubMed: 11466441]
- Li R, Hsieh CL, Young A, Zhang Z, Ren X, Zhao Z. Illumina synthetic long read sequencing allows recovery of missing sequences even in the "finished" *C. elegans* genome. *Sci Rep.* 2015; 5:10814.10.1038/srep10814 [PubMed: 26039588]
- McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet.* 2008; 40:1166–1174.10.1038/ng.238 [PubMed: 18776908]
- Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 2006; 16:1182–1190.10.1101/gr.4565806 [PubMed: 16902084]
- Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, et al. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics.* 2013; 194:987–996.10.1534/genetics.113.151381 [PubMed: 23792949]
- Monkley SJ, Pritchard CA, Critchley DR. Analysis of the mammalian talin2 gene TLN2. *Biochem Biophys Res Commun.* 2001; 286:880–885.10.1006/bbrc.2001.5497 [PubMed: 11527381]
- Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, et al. Genetic analysis of genome-wide variation in human gene expression. *Nature.* 2004; 430:743–747.10.1038/nature02797 [PubMed: 15269782]
- Murrell A, Heeson S, Cooper WN, Douglas E, Apostolidou S, Moore GE, et al. An association between variants in the IGF2 gene and Beckwith-Wiedemann syndrome: interaction between genotype and epigenotype. *Hum Mol Genet.* 2004; 13:247–255.10.1093/hmg/ddh013 [PubMed: 14645199]
- Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.* 2011; 7:e1001316.10.1371/journal.pgen.1001316 [PubMed: 21383968]
- Payseur BA, Jing P. A genomewide comparison of population structure at STRPs and nearby SNPs in humans. *Mol Biol Evol.* 2009; 26:1369–1377.10.1093/molbev/msp052 [PubMed: 19289600]
- Payseur BA, Jing P, Haasl RJ. A genomic portrait of human microsatellite variation. *Mol Biol Evol.* 2011; 28:303–312.10.1093/molbev/msq198 [PubMed: 20675409]

- Payseur BA, Place M, Weber JL. Linkage disequilibrium between STRPs and SNPs across the human genome. *Am J Hum Genet.* 2008; 82:1039–1050.10.1016/j.ajhg.2008.02.018 [PubMed: 18423524]
- Perry GH. The evolutionary significance of copy number variation in the human genome. *Cytogenet Genome Res.* 2008; 123:283–287.10.1159/000184719 [PubMed: 19287166]
- Pumpernik D, Oblak B, Borstnik B. Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Mol Genet Genomics.* 2008; 279:53–61.10.1007/s00438-007-0294-1 [PubMed: 17926066]
- Ram D, Leshkowitz D, Gonzalez D, Forer R, Levy I, Chowers M, et al. Evaluation of GS Junior and MiSeq next-generation sequencing technologies as an alternative to Trugene population sequencing in the clinical HIV laboratory. *J Virol Methods.* 2015; 212:12–16.10.1016/j.jviromet.2014.11.003 [PubMed: 25445792]
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001; 29:308–311. [PubMed: 11125122]
- Siegfried Z, Eden S, Mendelsohn M, Feng X, Tsuberi BZ, Cedar H. DNA methylation represses transcription in vivo. *Nat Genet.* 1999; 22:203–206.10.1038/9727 [PubMed: 10369268]
- Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet.* 2007; 39:226–231.10.1038/ng1955 [PubMed: 17206142]
- St George-Hyslop P, Haines J, Rogaev E, Mortilla M, Vaula G, Pericak-Vance M, et al. Genetic evidence for a novel familial Alzheimer's disease locus on chromosome 14. *Nat Genet.* 1992; 2:330–334.10.1038/ng1292-330 [PubMed: 1303289]
- Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature.* 2011; 480:490–495.10.1038/nature10716 [PubMed: 22170606]
- Stark AL, Hause RJ Jr, Gorsic LK, Antao NN, Wong SS, Chung SH, et al. Protein quantitative trait loci identify novel candidates modulating cellular response to chemotherapy. *PLoS Genet.* 2014; 10:e1004192.10.1371/journal.pgen.1004192 [PubMed: 24699359]
- Stein JL, Hua X, Morra JH, Lee S, Hibar DP, Ho AJ, et al. Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease. *Neuroimage.* 2010; 51:542–554.10.1016/j.neuroimage.2010.02.068 [PubMed: 20197096]
- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science.* 2007; 315:848–853.10.1126/science.1136678 [PubMed: 17289997]
- The International HapMap Consortium. The International HapMap Project. *Nature.* 2003; 426:789–796.10.1038/nature02168 [PubMed: 14685227]
- The International HapMap Consortium. A haplotype map of the human genome. *Nature.* 2005; 437:1299–1320.10.1038/nature04226 [PubMed: 16255080]
- The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007; 449:851–861.10.1038/nature06258 [PubMed: 17943122]
- The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010; 467:52–58.10.1038/nature09298 [PubMed: 20811451]
- The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* 2001; 409:860–921.10.1038/35057062 [PubMed: 11237011]
- Weber JL, Wong C. Mutation of human short tandem repeats. *Hum Mol Genet.* 1993; 2:1123–1128. [PubMed: 8401493]
- Westfall, P.; Young, S. Resampling-based multiple testing: examples and methods for p-value adjustment. New York: Wiley; 1993.
- Wooster R, Cleton-Jansen AM, Collins N, Mangion J, Cornelis RS, Cooper CS, et al. Instability of short tandem repeats (microsatellites) in human cancers. *Nat Genet.* 1994; 6:152–156.10.1038/ng0294-152 [PubMed: 8162069]
- Zhang DD, Cheng LJ, Badner JA, Chen C, Chen Q, Luo W, et al. Genetic control of individual differences in gene-specific methylation in human brain. *Am J Hum Genet.* 2010; 86:411–419.10.1016/j.ajhg.2010.02.005 [PubMed: 20215007]

- Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, et al. Evaluation of genetic variation contributing to differences in gene expression between populations. *Am J Hum Genet.* 2008; 82:631–640.10.1016/j.ajhg.2007.12.015 [PubMed: 18313023]
- Zhang W, Gamazon ER, Zhang X, Konkashbaev A, Liu C, Szilagy KL, et al. SCAN database: facilitating integrative analyses of cytosine modification and expression QTL. *Database (Oxford).* 2015; 201510.1093/database/bav025
- Zhang X, Cal AJ, Borevitz JO. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. *Genome Res.* 2011; 21:725–733.10.1101/gr.115337.110 [PubMed: 21467266]
- Zhang X, Moen EL, Liu C, Mu W, Gamazon ER, Delaney SM, et al. Linking the genetic architecture of cytosine modifications with human complex traits. *Hum Mol Genet.* 2014; 23:5893–905.10.1093/hmg/ddu313 [PubMed: 24943591]
- Zhang X, Mu W, Zhang W. On the analysis of the illumina 450k array data: probes ambiguously mapped to the human genome. *Front Genet.* 2012; 3:73.10.3389/fgene.2012.00073 [PubMed: 22586432]

Fig. 1. Identification of STRP-based mQTL

For both local (± 1 Mb windows) and whole-genome scans, linear models were used to identify STRP-based mQTL in CEU and YRI, separately: cytosine modification level $\sim$ length of STRPs + gender + error. *Cis*- and *trans*-acting STRP-based mQTL were then compared with SNP-based mQTL. Enrichment analysis for GWAS loci was performed for *cis*-acting STRP-based mQTL. KEGG pathway enrichment analysis was carried out for master STRPs (associated with > 200 CpGs) for *trans*-acting STRP-based mQTL.

Fig. 2. Cis-acting STRP-based mQTL are population-specific

(a) Venn diagram shows the number of STRP-based mQTL in CEU, YRI (p -value $< 10^{-3}$), and the previously detected SNP-based mQTL; (b) Venn diagram shows the number of STRP-based mQTL in CEU, YRI (adjusted p -value < 0.05 , corresponding to nominal p -value $< 4.0 \times 10^{-5}$), and the previously detected SNP-based mQTL. (c) Population specificity of STRP-based mQTL (p -value $< 10^{-3}$; X-axis: r^2 of YRI, Y-axis: r^2 of CEU); (d) Population specificity of STRP-based mQTL (adjusted p -value < 0.05 , corresponding to nominal p -

value $< 4.0 \times 10^{-5}$; X- axis: r^2 of YRI, Y axis: r^2 of CEU). SNP-based mQTL were identified using the SCAN database with a p -value $< 10^{-4}$.

Fig. 3. Chromosomal distributions of STRP-based mQTL

The heatmap shows chromosomal distributions of all STRP-based mQTL from the whole-genome scans in (a) CEU; and (b) YRI. Besides *cis*-acting relationships, STRP-based mQTL are predominantly *trans*-acting.

Fig. 4. Enrichment of GWAS loci among *cis*-acting STRP-based mQTL

The null distributions of the numbers of SNPs overlapped with GWAS loci are displayed as histograms. The asterisk marks the true number of SNPs overlapped with GWAS loci within different windows: **(a)** $\pm 100\text{Kb}$; **(b)** $\pm 500\text{Kb}$; and **(c)** $\pm 1\text{Mb}$, of *cis*-acting STRP-based mQTL (adjusted p -value < 0.05).

Table 1

Summary of local and genome-wide scan

Type of scan	Cutoff	Population	No. of associations	% associations	No. of CpGs ^a	% CpGs ^b	Total associations ^c	Total CpGs ^c
Local scan	p -value < 10^{-3}	CEU	183	0.12	183	0.13	471	471
		YRI	173	0.11	173	0.13		
		Combined	176	0.11	176	0.13		
Genome-wide scan	p -value < 10^{-5}	CEU	21	0.01	21	0.02	43	43
		YRI	9	0.01	9	0.01		
		Combined	21	0.01	21	0.02		
		CEU	139420	0.07	41794	14.74	259108	74676
		YRI	76817	0.04	30994	10.93		
		Combined	121866	0.06	50104	17.67		

^aThe number of unique CpG sites mapped.

^bThe percentage of unique CpG sites mapped among the unique CpG sites analyzed.

^cUnion of associations and CpGs among CEU and YRI samples.