Title:

Implicating candidate genes at GWAS signals by leveraging topologically associating domains

Running Title:

TADs aid candidate gene discovery from GWAS

Authors:

Gregory P. Way[1,2], Daniel W. Youngstrom[3], Kurt D. Hankenson[3], Casey S. Greene[2]*, and Struan F.A. Grant[4,5]*

* These authors directed this work jointly

Affiliations:

[1]Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, PA 19104, USA

[2]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]Department of Orthopaedic Surgery, School of Medicine, University of Michigan, Ann Arbor, MI, USA

[4]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[5]Division of Human Genetics, Division of Endocrinology and Diabetes, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

Co-Corresponding Authors:*

Struan F.A. Grant

Rm 1102D, 3615 Civic Center Blvd
Philadelphia, PA 19104
Office: 267-426-2795
Fax: 215-590-1258

Casey S. Greene

10-131 SCTR 34th and Civic Center Blvd,
Philadelphia, PA 19104
Office: 215-573-2991
Fax: 215-573-9135

Conflict of Interest:

The authors have nothing to disclose.

Sources of Support:

This work was supported by the Genomics and Computational Biology Graduate program at The University of Pennsylvania (to G.P.W.); the Gordon and Betty Moore Foundation's Data Driven Discovery Initiative (grant number GBMF 4552 to C.S.G); the National Institute of Dental & Craniofacial Research (grant number NIH F32DE026346 to D.W.Y.); S.F.A.G is supported by the Daniel B. Burke Endowed Chair for Diabetes Research. D.W.Y. is supported by NIH-NRSA F32 DE

Abstract:

Genome wide association studies (GWAS) have contributed significantly to the understanding of complex disease genetics. However, GWAS only report associated signals and do not necessarily identify culprit genes. As most signals occur in non-coding regions of the genome, it is often challenging to assign genomic variants to the underlying causal mechanism(s). Topologically associating domains (TADs) are primarily cell-type independent genomic regions that define interactome boundaries and can aid in the designation of limits within which an association most likely impacts gene function. We describe and validate a computational method that uses the genic content of TADs to discover candidate genes. Our method, called "TAD_Pathways", performs a Gene Ontology (GO) analysis over genes that reside within TAD boundaries corresponding to GWAS signals for a given trait or disease. We applied our pipeline to the GWAS catalog entries associated with bone mineral density (BMD), identifying 'Skeletal System Development' (Benjamini-Hochberg adjusted $p=1.02\text{x}10^{-5}$) as the top ranked pathway. In many cases, our method implicated a gene other than the nearest gene. Our molecular experiments describe a novel example: *ACP2*, implicated at the canonical '*ARHGAP1*' locus. We found *ACP2* to be an important regulator of osteoblast metabolism, whereas *ARHGAP1* was not supported. Our results via the example of BMD demonstrate how basic principles of three-dimensional genome organization can define biologically informed association windows.

Keywords:

Candidate gene method, genome wide association study, topologically associating domains, pathway analysis, bone mineral density

Introduction:

GWAS have been applied to over 300 different traits, leading to the discovery of important disease associations[1]. However, assigning signals to causal genes has proven difficult because these signals fall principally within noncoding regions and do not necessarily implicate the nearest gene[2,3]. For example, a signal found in an *FTO* intron has been shown to physically interact with and lead to the differential expression of other genes, and not *FTO* itself [4]. Moreover, there is evidence suggesting a type 2 diabetes GWAS association impacting *TCF7L2* also influences *ACSL5* [5]. It remains unclear how pervasive these kinds of associations are, but similar strategies are necessary in order for GWAS to better guide research and precision medicine[6].

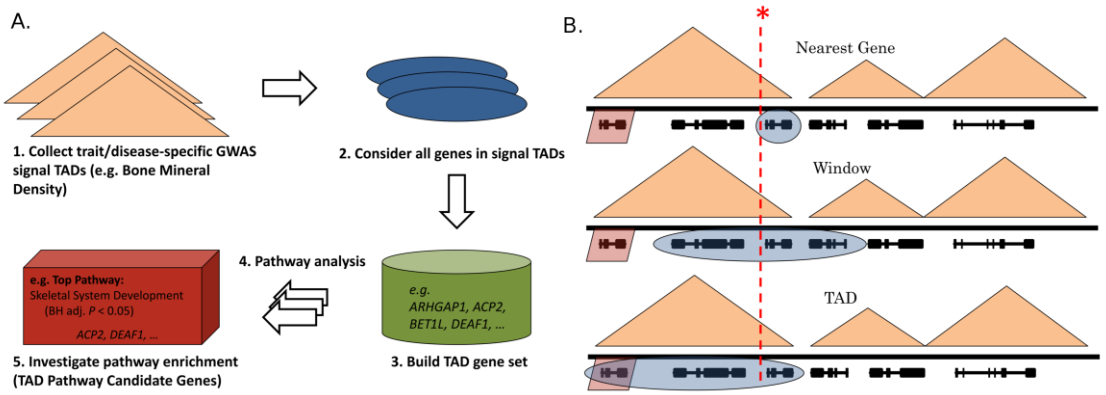
Chromatin interaction studies have discovered key genome organization principles including TADs[7]. TADs are genomic regions defined by increased contact frequency, consistency across cell types, and their enrichment of insulator element flanks[8]. Therefore, TADs can be used as boundaries of where non-coding causal variants will most likely impact tissue independent function.

We developed a computational approach, called "TAD_Pathways", which uses TADs to determine candidate genes. We validated our method using BMD GWAS[9]. Using our method, we identified *ACP2* as a novel regulator of osteoblast metabolism.

Methods:

Computational procedures to identify candidate genes

TAD_Pathways is a computational method that uses TADs to discover candidate genes (**Figure 1A**). Alternative approaches either assign genes based on nearest gene or by an arbitrary or a linkage disequilibrium-based window of several kilobases[10] (**Figure 1B**). Here, we use human embryonic stem cell TAD boundaries as reported by Dixon *et al.* and converted to hg19 by Ho *et al.* to build TAD based genesets that consists of all Gencode genes that fall inside TADs implicated with BMD associations[8,11]. We perform a pathway overrepresentation test[12] for the input TAD genes against Gene Ontology (GO) terms[13]. This determines if the geneset is associated with any term at a higher probability than by chance. We included both experimentally confirmed and computationally inferred genes, which permit the inclusion of putative genes which do not necessarily have literature support, but are predicted computationally. For validation, we consider only the most significantly enriched term, but a user can also select multiple. Our method also supports custom input SNP lists. TAD_Pathways software is available at https://github.com/greenelab/tad_pathways.



106

107 **Figure 1.** *Concepts motivating our approach.* Topologically associating domains (TADS)
108 are shown as orange triangles, genes are shown as black lines, and a genome wide
109 significant GWAS signal is shown as a dotted red line. **(A)** The TAD_Pathways method.
110 An example using Bone Mineral Density GWAS signals is shown. **(B)** Three hypothetical
111 examples illustrated by a cartoon. The ground truth causal gene is shaded in red. The
112 method-specific selected genes are shaded in blue. The top panel describes a nearest gene
113 approach. The nearest gene in this scenario is not the gene actually impacted by the GWAS
114 SNP. The middle panel describes a window approach. Based either on linkage
115 disequilibrium or an arbitrarily sized window, the scenario does not capture the true gene.
116 The bottom panel describes the TAD_Pathways approach. In this scenario, the causal gene
117 is selected for downstream assessment.
118

119 *Experimental knockdown of candidate genes*

120 We investigated two candidate genes predicted by TAD_Pathways; *ACP2* and *DEAF1*.
121 The corresponding BMD GWAS loci rs7932354 (cytoband: 11p11.2) and rs11602954
122 (cytoband: 11p15.5) are assigned to *ARHGAP1* and *BET1L*, respectively. The candidate
123 genes were selected because no experimental support exists for their impact in bone and the
124 nearest genes to the GWAS signals do not have immediate bone related functional
125 implications. All experiments were conducted in three temporally separated independent
126 technical replicates from cryopreserved P2 aliquots of a human fetal osteoblast cell line
127 (ATCC hFOB 1.19 CRL-11372). Full experimental design can be found in the
128 accompanying **Supplementary Information**.

siRNA transfections were conducted using a commercial reagent system with cells assigned to one of 8 experimental groups: untreated control, siRNA-negative transfection reagent control, scrambled control siRNA, *TNAP* siRNA, *ARHGAP1* siRNA, *ACP2* siRNA, *BET1L* siRNA or Suppressin (*DEAF1*) siRNA.

For quantitative gene expression analysis, whole RNA was harvested at Day 4, isolated and reverse-transcribed. Duplicate 20ng templates for each gene of interest were assayed via real-time PCR using SYBR reagents. Results were analyzed using the 2^{-ddCt} method using GAPDH as a housekeeping gene and reported as (mean±standard deviation) fold-change versus untreated controls. Cellular metabolism/proliferation was assessed at 1 and 4 days post-transfection using an MTT cell growth determination kit. Results are reported as the difference in mean absorbance at 570nm minus 690nm from Day 1 to Day 4 and error bars represent root-mean-square standard deviation from the measurements at both days. Early osteoblast differentiation was assessed at 4 days post-transfection using an alkaline phosphatase (ALP) kit. Dried plates were scanned and ALP+ intensities/areas were quantified in ImageJ. Values are reported as mean ± standard deviation.

Results:

TAD_Pathways reveals candidate genes within phenotype-associated TADs

We applied TAD_Pathways to BMD GWAS results derived from replication-requiring journals[9,14–16]. Our method implicated 'Skeletal System Development' as the top ranked pathway (Benjamini-Hochberg adjusted $p=1.02\text{x}10^{-5}$). For full BMD TAD_Pathways refer to **Supplementary Table S1**. Many candidates were not the nearest gene to the GWAS signal and several had independent eQTL support (**Supplementary Table S2**).

siRNA Knockdown of TAD_Pathways Gene Predictions in Osteoblast Cells

We targeted the expression of four of these genes *in vitro* using siRNA and assessed transcriptional knockdown efficiency. Knockdown efficiencies were: *TNAP* 48.7±9.9% (p=0.141), *ARHGAP1* 68.7±14.3% (p=0.015), *ACP2* 48.9±6.4% (p=0.035), *BET1L* 56.4±1.0% (p>0.05) and *DEAF1* 52.7±9.2% (p=0.021) (**Figure 2**).

We noted variation across the three controls, with the scrambled siRNA control altering expression of *OCN* (osteocalcin), *IBSP* (bone sialoprotein), *TNAP* and *BET1L* (p<0.05). Relative to the scrambled siRNA control, *OCN* was downregulated in all siRNA groups (p<0.05) except for *BET1L* siRNA (p=0.122). *OSX*, *IBSP* and *TNAP* were not significantly altered by any siRNA treatment (**Figure 2**).

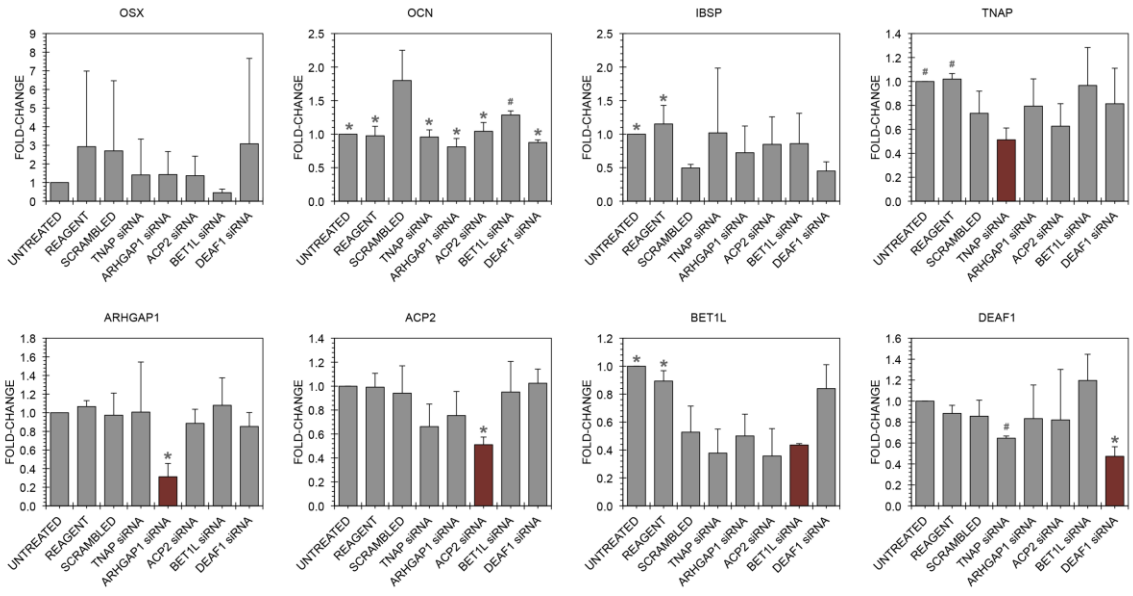


Figure 2: *Real-time PCR of osteoblast differentiation genes and GWAS/TAD hits in hFOB cells.* siRNA was used to knock down expression of *TNAP* (positive control), *ARHGAP1*, *ACP2*, *BET1L* and *DEAF1*. Relative expression of the osteoblast marker genes *OSX*, *OCN* and *IBSP* suggest that GWAS/TAD hits are not major regulators of bone differentiation in this model. Red bars highlight specificity of each siRNA knockdown. Values represent mean ± standard deviation. Statistical significance relative to the scrambled siRNA control is annotated as: *$p \leq 0.05$ and #$p \leq 0.10$ using a two-tailed Student's *t*-test.

Metabolic Activity of TAD_Pathways Gene Predictions

Treatment with *ACP2* siRNA led to a 66.0% reduction in MTT metabolic activity versus the scrambled siRNA control (p=0.012). *ARHGAP1* siRNA caused a 38.8% reduction, falling short of statistical significance (p=0.088). siRNA targeted against *TNAP*, *BET1L* or *DEAF1* did not alter MTT metabolic activity (**Figure 3A**).

Alkaline Phosphatase Activity of TAD_Pathways Gene Predictions

ALP is highly expressed in osteoblasts; disruption of proliferation or osteoblast differentiation would result in downregulation of ALP. *TNAP* siRNA significantly reduced ALP intensity by 5.98±1.77 units versus the scrambled siRNA control (p=0.006). *ACP2* siRNA also significantly reduced ALP intensity by 8.74±2.11 (p=0.003). The control stained less intensely than untreated or transfection reagent controls, but this did not reach statistical significance ($0.05<p<0.10$) (**Figure 3B**).

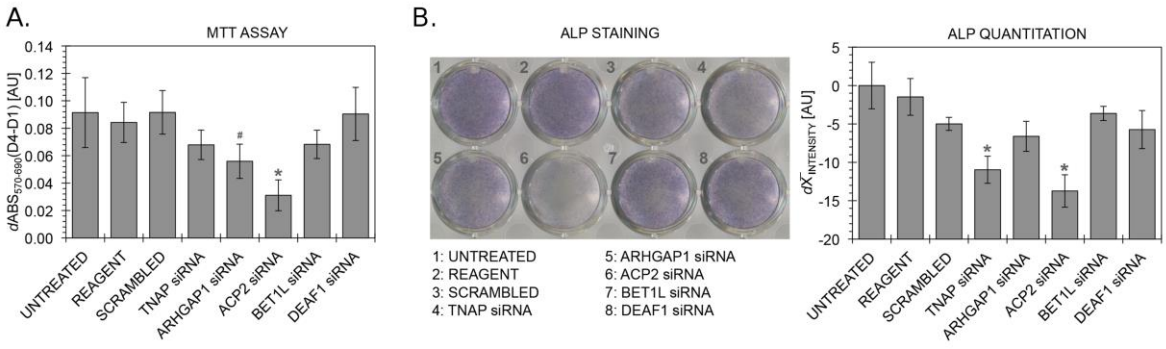


Figure 3. *Validating two 'TAD Pathway' predictions for Bone Mineral Density GWAS hits on hFOB cells.* siRNA was used to knock down expression of *TNAP*, *ARHGAP1*, *ACP2*, *BET1L* and *DEAF1*. **(A)** Knockdown of *ACP2* decreases cellular metabolic activity, demonstrated using an MTT assay. **(B)** ALP staining and quantitation indicates that knockdown of *TNAP* or *ACP2* inhibits performance in an osteoblast differentiation assay. Values represent mean ± standard deviation. Statistical significance relative to the scrambled siRNA control is annotated as: *$p \leq 0.05$ and #$p \leq 0.10$ using a two-tailed Student's *t*-test.

Discussion:

We showed that TAD_Pathways can reveal functional gene to intermediate phenotype relationships using BMD. Several of the TAD_Pathways genes, such as *LRP5*, are *bona fide* BMD genes already identified by nearest gene GWAS, eQTL analyses, and clinical syndromes[17], thus providing positive controls. However, several BMD GWAS signals do not have obvious nearest-gene associations with bone. Our results suggest that a nearby gene *ACP2,* and not the nearest gene *ARHGAP1*, regulates osteoblast proliferation/viability.

There are several limitations to the approach. Publication biases from pathway curation present challenges[18]. To lessen this bias, we include curated and computationally predicted GO annotations. Additionally, we used TAD boundaries defined by Dixon *et al.*, whereas increased Hi-C resolution reduced estimated TAD sizes[19]. Despite our method using larger TADs, thus including more presumably false positive genes, we still identify relevant pathways. Furthermore, the method will likely fail in a disease instigated by aberrant looping. We were also concerned that TAD_Pathways would work only in BMD. Indeed, when we applied TAD_Pathways to Type 2 Diabetes, we also identify several candidate genes that are not the nearest gene (see Supplementary Table S3). Moreover, the experimental validation was performed in a simplified *in vitro* cell culture system lacking organismal complexity, and the cell line selected is tetraploid, which may partially compensate for gene knockdown. While TAD_Pathways identified several candidate genes, we only examined two, and our validation approach does not directly interrogate each SNP. Lastly, one of the investigated GWAS SNPs, rs7932354, located in the *ARHGAP1* promoter, is an *ARHGAP1* eQTL in several tissues as determined by GTEx[20]. However,

none of these tissues are bone related and our screen did not implicate *ARHGAP1* in osteoblast processes.

In conclusion, TAD_Pathways can be used as a candidate gene discovery tool that uses chromatin looping to map associations to candidate genes. We also validated our method by implicating *ACP2* as a gene involved in BMD determination, which warrants future investigation. We believe TAD_Pathways and algorithms that leverage 3D genomic structure will be poised to discover novel disease features.

Acknowledgements

Hannah E. Sexton and Troy L. Mitchell assisted in optimizing siRNA transfection conditions. Daniel Himmelstein and Amy Campbell performed analytical code review.

Conflict of Interest

The authors have nothing to disclose.

Availability of data and material

All data used to construct the TAD_Pathways approach are publically available datasets. We make all software used to develop this approach publically available in a GitHub repository (http://github.com/greenelab/tad_pathways). We also provide a docker image (https://hub.docker.com/r/gregway/tad_pathways/) and archive the GitHub software on Zenodo (https://zenodo.org/record/163950).

References:

1 Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; **42**: D1001–D1006.

2 Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? *Nucleic Acids Res* 2016; : gkw500.

3 Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 2012; **30**: 1095–1106.

4 Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C *et al. FTO* Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* 2015; **373**: 895–907.

5 Xia Q, Chesi A, Manduchi E, Johnston BT, Lu S, Leonard ME *et al.* The type 2 diabetes presumed causal variant within TCF7L2 resides in an element that controls the expression of ACSL5. *Diabetologia* 2016; **59**: 2360–2368.

6 Herman MA, Rosen ED. Making Biological Sense of GWAS Data: Lessons from the FTO Locus. *Cell Metab* 2015; **22**: 538–539.

7 Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 2009; **326**: 289–293.

8 Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012; **485**: 376–380.

9 Estrada K, Styrkarsdottir U, Evangelou E, Hsu Y-H, Duncan EL, Ntzani EE *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet* 2012; **44**: 491–501.

10 Hao K, Di X, Cawley S. LdCompare: rapid computation of single- and multiple-marker r2 and genetic coverage. *Bioinformatics* 2007; **23**: 252–254.

11 Ho JWK, Jung YL, Liu T, Alver BH, Lee S, Ikegami K *et al.* Comparative analysis of metazoan chromatin organization. *Nature* 2014; **512**: 449–452.

12 Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* 2013; **41**: W77–W83.

13 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* 2000; **25**: 25–29.

14 Richards JB, Rivadeneira F, Inouye M, Pastinen TM, Soranzo N, Wilson SG *et al.* Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet Lond Engl* 2008; **371**: 1505–1512.

15 Rivadeneira F, Styrkársdottir U, Estrada K, Halldórsson BV, Hsu Y-H, Richards JB *et al.* Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet* 2009; **41**: 1199–1206.

16 Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, Jonasdottir A *et al.* Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. *Nature* 2013; **497**: 517–520.

17 Baron R, Kneissel M. WNT signaling in bone homeostasis and disease: from human mutations to treatments. *Nat Med* 2013; **19**: 179–192.

18 Greene CS, Troyanskaya OG. Accurate evaluation and analysis of functional genomics data and methods: Accurate evaluation and analysis of functional genomics data and methods. *Ann N Y Acad Sci* 2012; **1260**: 95–100.

19 Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 2014; **159**: 1665–1680.

20 Aguet F, Brown AA, Castel S, Davis JR, Mohammadi P, Segre AV *et al.* Local genetic effects on gene expression across 44 human tissues. 2016http://biorxiv.org/lookup/doi/10.1101/074450 (accessed 3 Jan2017).