1   **Title:**

2   Determining causal genes from GWAS signals using topologically associating domains

3

4   **Authors:**

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

6   Struan F.A. Grant[4,5]*

7   * These authors directed this work jointly

8

9   **Affiliations:**

10   [1]Genomics and Computational Biology Graduate Program, University of Pennsylvania,

11   Philadelphia, PA 19104, USA

12   [2]Department of Systems Pharmacology and Translational Therapeutics, University of

13   Pennsylvania, Philadelphia, PA 19104, USA

14   [3]Department of Small Animal Clinical Sciences, College of Veterinary Medicine, Michigan

15   State University, East Lansing, MI 48824, USA

16   [4]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania,

17   Philadelphia, PA 19104, USA

18   [5]Division of Human Genetics, Division of Endocrinology and Diabetes, The Children's

19   Hospital of Philadelphia, Philadelphia, PA 19104, USA

20

21   **Author Emails:**

22   GPW: gregway@mail.med.upenn.edu; DWY: dwy@msu.edu; KDH: kdhank@msu.edu;

23   CSG: csgreene@mail.med.upenn.edu; *SFAG: grants@email.chop.edu

24   **Co-Corresponding Authors:***

Struan F.A. Grant   Casey S. Greene

Rm 1102D, 3615 Civic Center Blvd   10-131 SCTR 34th and Civic Center Blvd,

Philadelphia, PA 19104   Philadelphia, PA 19104

Office: 267-426-2795   Office: 215-573-2991

Fax: 215-590-1258   Fax: 215-573-9135

25

26   **Abstract:**

27   *Background*

28   Genome wide association studies (GWAS) have contributed significantly to the field of

29   complex disease genetics. However, GWAS only report signals associated with a given trait and

30   do not necessarily identify the precise location of culprit genes. As most association signals

31   occur in non-coding regions of the genome, it is often challenging to assign genomic variants to

32   the underlying causal mechanism(s). Topologically associating domains (TADs) are primarily

33   cell-type independent genomic regions that define interactome boundaries and can aid in the

34   designation of limits within which a GWAS locus most likely impacts gene function.

35   *Results*

36   We describe and validate a computational method that uses the genic content of TADs to

37   assign GWAS signals to likely causal genes. Our method, called "TAD_Pathways", performs a

38   Gene Ontology (GO) analysis over all genes that reside within the boundaries of all TADs

39   corresponding to the GWAS signals for a given trait or disease. We applied our pipeline to the

40   GWAS catalog entries associated with bone mineral density (BMD), identifying 'Skeletal

41   System Development' (Benjamini-Hochberg adjusted $p = 1.02\text{x}10^{-5}$) as the top ranked pathway.

42   Often, the causal gene identified at a given locus was well known and/or the nearest gene to the

43   sentinel SNP. In other cases, our method implicated a gene further away. Our molecular

44   experiments describe a novel example: *ACP2*, implicated at the canonical '*ARHGAP1*' locus. We

45   found *ACP2* to be an important regulator of osteoblast metabolism, whereas a causal role of

46   *ARHGAP1* was not supported.

47   *Conclusions*

48   Our results demonstrate how basic principles of three-dimensional genome organization can

49   help define biologically informed windows of signal association. We anticipate that

50   incorporating TADs will aid in refining and improving the performance of a variety of

51   algorithms that linearly interpret genomic content.

52

53   **Keywords:**

54   Genome wide association study, gene prioritization, topologically associating domains,

55   pathway analysis, bone mineral density

56

57   **Background:**

58   Genome-wide association studies (GWAS) have been applied to over 300 different traits,

59   leading to the discovery and subsequent validation of several important disease associations [1].

60   However, GWAS can only discover association signals in the data. Subsequent assignment of

61   signal to causal genes has proven difficult due to these signals falling principally within

62   noncoding genomic regions [2–4] and not necessarily implicating the nearest gene [5]. For

63   example, a signal found within an intron for *FTO*, a well-studied gene previously thought to be

64   important for obesity [6], has been shown to physically interact with and lead to the differential

65   expression of two genes (*IRX3* and *IRX5*) directly next to this gene, and not *FTO* itself [7–9].

66   Moreover, there is evidence suggesting a type 2 diabetes GWAS association previously

67   implicating *TCF7L2* [10] also influences the nearby *ACSL5* gene [11]. It remains unclear how

68   pervasive these kinds of associations are, but similar strategies are necessary in order for GWAS

69   to better guide research and precision medicine [12].

70   Three-dimensional genomics has changed the way geneticists think about genome

71   organization and its functional implications [13,14]. Genome-wide chromatin interaction maps

72   have facilitated the development of several genome organization principles, including

73   topologically associating domains (TADs) [15–18]. TADs are sub-architectural units of the

74   overall genome organization that have consistent and functionally important genomic element

75   distributions including an enrichment of housekeeping genes, insulator elements, and early

76   replication timing regions at boundary regions [19–21]. TADs are largely consistent across

77   different cell types and demonstrate synteny [22,23]. These observations can therefore allow the

78   leveraging of TADs to set the bounds of where non-coding causal variants can most likely

79   impact promoters, enhancers and genes in a tissue independent fashion [24,25]. Therefore, we

80   sought to develop a method that integrates GWAS data with interactome boundaries to more

81   accurately map signals to the mostly likely candidate gene(s).

82   We developed a computational approach, called "TAD_Pathways", which is agnostic to gene

83   locations relative to each GWAS signal within TADs. We scanned publically available GWAS

84   data for given traits and used TAD boundaries to output lists of genes likely to be causal. We

85   demonstrate this approach by assessing the influence of GWAS signals on bone mineral density

86   (BMD) [26–29]. This trait is clinically of great importance as low BMD is an important

87   precursor to osteoporosis, a disease condition affecting millions of patients annually [30]. We

88   also chose BMD as a trait for analysis because BMD GWAS primarily points to very well-

89   known genes involved in bone development (positive controls) but there remain a number of

90   established loci where no obvious gene resides, therefore offering the opportunity to uncover

91 novel biology. After applying our TAD_Pathways discovery approach, we investigated putative

92 causal genes using cell culture-based assays, identifying *ACP2* as a novel regulator of osteoblast

93 metabolism.

94

95 **Results:**

96 *The Genomic Landscape of SNPs across Topologically Associating Domains*

97 We observed a consistent and non-random distribution of SNPs across TADs derived from

98 human embryonic stem cells (hESC), human fibroblasts (IMR90), mouse embryonic stem cells

99 (mESC), and mouse cortex cells (mcortex) cells. As expected, SNPs are tightly associated with

100 TAD length for each cell type, but there are substantial outlier TADs (**Figure 1**). For example,

101 the TAD harboring the largest number of common SNPs (minor allele frequency (MAF) greater

102 than 0.05) in hESC is located on chromosome 6 (UCSC hg19: chr6:31492021-32932022) and

103 has 19,431 SNPs. Not surprisingly, this TAD harbors an abundance of genes including *HLA*

104 genes, which are well known to have many polymorphic sites [31]. However, the other human

105 cell line (IMR90) outlier TAD is located on chromosome 8 (UCSC hg19: chr8:2132593-

106 6252592) and has 27,220 SNPs and could be potentially biologically meaningful. Indeed,

107 although this TAD harbors relatively few genes, it does include *CSMD1*, a gene implicated in

108 cancer and neurological disorders such as epilepsy and schizophrenia [32,33].

109 Common SNPs were enriched near the center of TADs (**Figure 1A**). This is the opposite of

110 gene (**Supplementary Figure S1**) and repeat element (**Supplementary Figure S2**) distributions

111 (also see Dixon *et al.* 2012) [22]. The repeat element distribution was driven largely by the

112 SINE/Alu repeat distribution, which could not be explained by GC content and estimated

113 evolutionary divergence (**Supplementary Figure S3**). We also observed that common SNPs are

114   significantly enriched in the 3' half of TADs in hESC and mcortex cells (**Supplementary Table**

115   **S1**). There was also a slight increase in GWAS implicated SNPs near hESC TAD boundaries

116   (**Figure 1B**). Given the non-random patterns observed across the TADs, we went on to explore

117   the gene content further in an attempt to imply causality at given GWAS loci.

118
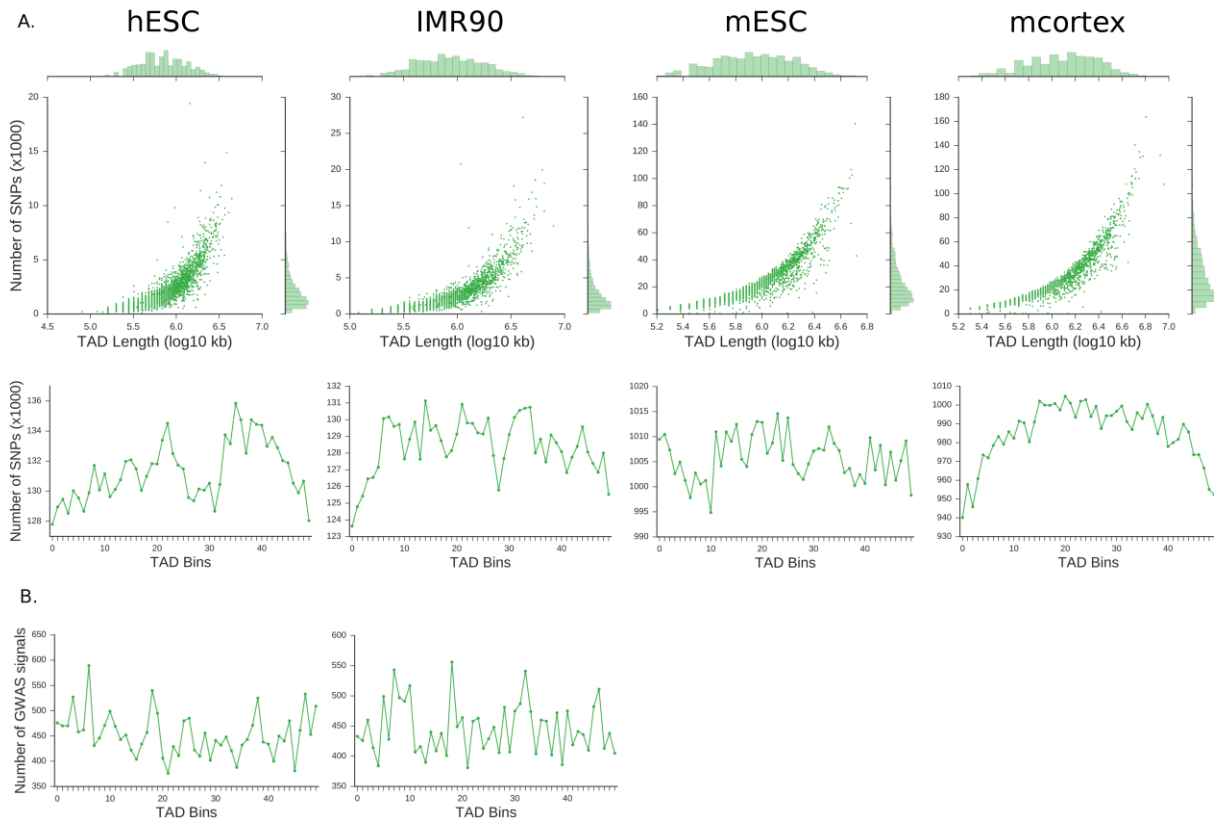
119

120   **Figure 1.** *Distributions of single nucleotide polymorphisms (SNPs) across topologically*
121   *associating domains of four cell types.* (**A**) Top – The length of TADs is associated with the
122   number of SNPs found in each cell type. Bars on the top and right side of the plots represent
123   histograms of each respective metric. hESC and IMR90 TADs are based on 1000 genomes phase
124   3 SNPs (hg19) and mESC and mcortex TADs represent 15 strains from the mouse genomes
125   project version 2 (mm9). Bottom – Number of SNPs found in each cell type. (**B**) Number of
126   GWAS SNPs from the NHGRI-EBI GWAS catalog that reached genome-wide significance in
127   replication required journals. In each case, we independently discretized TADs into 50 bins
128   where the distribution of elements is linear from 5' to 3' (bin 0 is the 5' most end of all TADs).
129

130

6

131   *TAD_Pathways reveals potentially causal genes within phenotype-associated TADs*

132   Seeking to leverage TADs and disease associated SNPs, we integrated GWAS and TAD

133   domain boundaries in an effort to assign GWAS signals to causal genes. Alternative approaches

134   to understand the gene landscape of a locus that do not consider TAD boundaries typically either

135   assign genes to a GWAS signal based on nearest gene [34] or by an arbitrary or a linkage

136   disequilibrium-based window of several kilobases [35,36] (**Figure 2A**). Instead, we used TAD

137   boundaries and the full catalog of GWAS findings for a given trait or disease to assign genes to

138   GWAS variants based on overrepresentation in a gene set in an approach we termed

139   "TAD_Pathways". For a given trait or disease, we collected all genes that are located in TADs

140   harboring significant GWAS signals. We then applied a statistical enrichment analysis for

141   biological pathways using this TAD gene set and assign candidate genes within a TAD based on

142   the pathways significantly associated with a phenotype (**Figure 2B**). In our implementation, we

143   used GO biological processes, GO cellular components, and GO molecular functions to provide

144   the pathway sets [37]. We included both experimentally confirmed and computationally inferred

145   GO gene annotations, which permit the inclusion of putative casual genes that do not necessarily

146   have literature support but are predicted by a variety of computational methods.

147   To validate our approach, we applied TAD_Pathways to bone mineral density (BMD)

148   GWAS results derived from replication-requiring journals [26–29]. Our method implicated

149   'Skeletal System Development' (Benjamini-Hochberg adjusted $p = 1.02 \times 10^{-5}$) as the top ranked

150   pathway. We provide full TAD_Pathways results for BMD in **Supplementary Table S2**.

151   Despite a high content of presumably non-causal genes, which we expect would contribute noise

152   to the overrepresentation analysis [38], our method demonstrated enrichment of a skeletal system

153   related pathway and selected a subset of potentially causal genes belonging to the same pathway.

7

154   Many of these genes (24/38) were not the nearest gene to the GWAS signal and several also had

155   independent expression quantitative trail loci (eQTL) support (**Supplementary Table S3,**

156   **Supplementary Figure S4**).

157
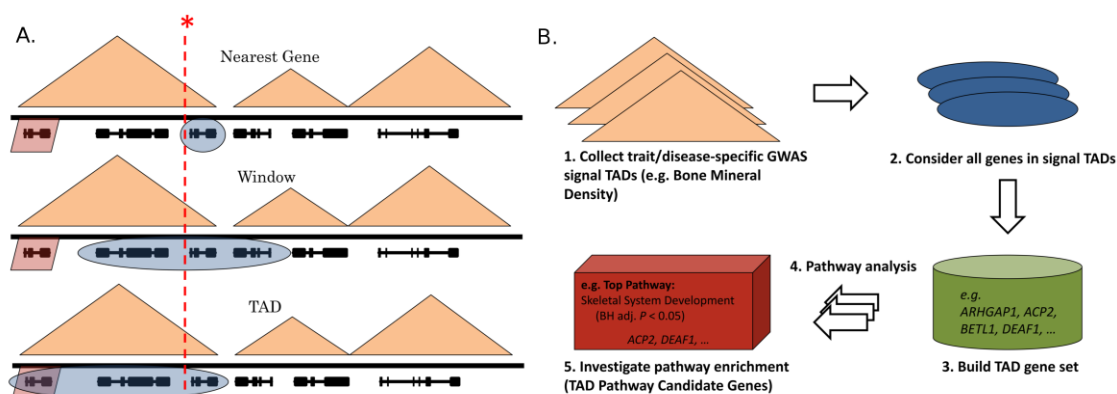
158

159   **Figure 2.** *Concepts motivating our approach.* Topologically associating domains (TADS) are
160   shown as orange triangles, genes are shown as black lines, and a genome wide significant
161   GWAS signal is shown as a dotted red line. **(A)** Three hypothetical examples illustrated by a
162   cartoon. The ground truth causal gene is shaded in red. The method-specific selected genes are
163   shaded in blue. The top panel describes a nearest gene approach. The nearest gene in this
164   scenario is not the gene actually impacted by the GWAS SNP. The middle panel describes a
165   window approach. Based either on linkage disequilibrium or an arbitrarily sized window, the
166   scenario does not capture the true gene. The bottom panel describes the TAD_Pathways
167   approach. In this scenario, the causal gene is selected for downstream assessment. **(B)** The
168   TAD_Pathways method. An example using Bone Mineral Density GWAS signals is shown.
169

170   *siRNA Knockdown of TAD Pathway Gene Predictions in Osteoblast Cells*

171   The loci rs7932354 (cytoband: 11p11.2) and rs11602954 (cytoband: 11p15.5) are currently

172   assigned to *ARHGAP1* and *BETL1* but our method implicated *ACP2* and *DEAF1*, respectively.

173   The two genes implicated by TAD_Pathways, *ACP2* and *DEAF1*, lacked eQTL support and were

174   not the nearest gene to the BMD GWAS signal. We tested the gene expression activity and

175   metabolic importance of these four genes, *ARHGAP1*, *BETL1*, *DEAF1*, and *ACP2*. Specifically,

176   our assays in a human fetal osteoblast cell line (hFOB) evaluate whether or not the

8

177   TAD_Pathways method identifies causal genes at GWAS signals beyond those captured by

178   closest and eQTL connected genes. Though these two genes were annotated to the identified GO

179   process, the annotation had been made computationally and their known biology did not provide

180   obvious links to bone biology.

181   We targeted the expression of all four of these genes *in vitro* using small interfering RNA

182   (siRNA), assessing knockdown efficiency at the mRNA level relative to untreated controls and

183   determined corresponding *p* values relative to scrambled siRNA controls. We used an siRNA

184   targeting tissue-nonspecific alkaline phosphatase (*TNAP*) as a positive control. Knockdown

185   efficiencies were: *TNAP* siRNA 48.7±9.9% (*p*=0.141), *ARHGAP1* siRNA 68.7±14.3%

186   (*p*=0.015), *ACP2* siRNA 48.9±6.4% (*p*=0.035), *BET1L* 56.4±1.0% (N.S.) and *DEAF1*

187   52.7±9.2% (*p*=0.021) (**Figure 3**). siRNA targeted against each gene of interest did not down-

188   regulate the expression of the other genes under investigation, indicating specificity of

189   knockdown, although we noted that *TNAP* siRNA did reduce *DEAF1* gene expression, though

190   this did not reach the threshold for statistical significance (*p*= 0.077).

191   We noted significant variation across the three controls, with the scrambled siRNA control

192   altering expression of *OCN* (osteocalcin), *IBSP* (bone sialoprotein), *TNAP* and *BET1L* ($p < 0.05$).

193   Relative to the scrambled siRNA control, *OCN* was downregulated in all siRNA groups ($p <$

194   0.05) except for *BET1L* siRNA ($p = 0.122$). *OSX*, *IBSP* and *TNAP* were not significantly altered

195   by any siRNA treatment (**Figure 3**).

196

197   *Metabolic Activity of TAD Pathway Gene Predictions*

198   Use of *ACP2* siRNA led to a 66.0% reduction in MTT metabolic activity versus the

199   scrambled siRNA control ($p = 0.012$). *ARHGAP1* siRNA caused a 38.8% reduction, which fell

200   short of statistical significance (*p* = 0.088). siRNA targeted against *TNAP*, *BET1L* or *DEAF1* did

201   not alter MTT metabolic activity (**Figure 4A**).

202
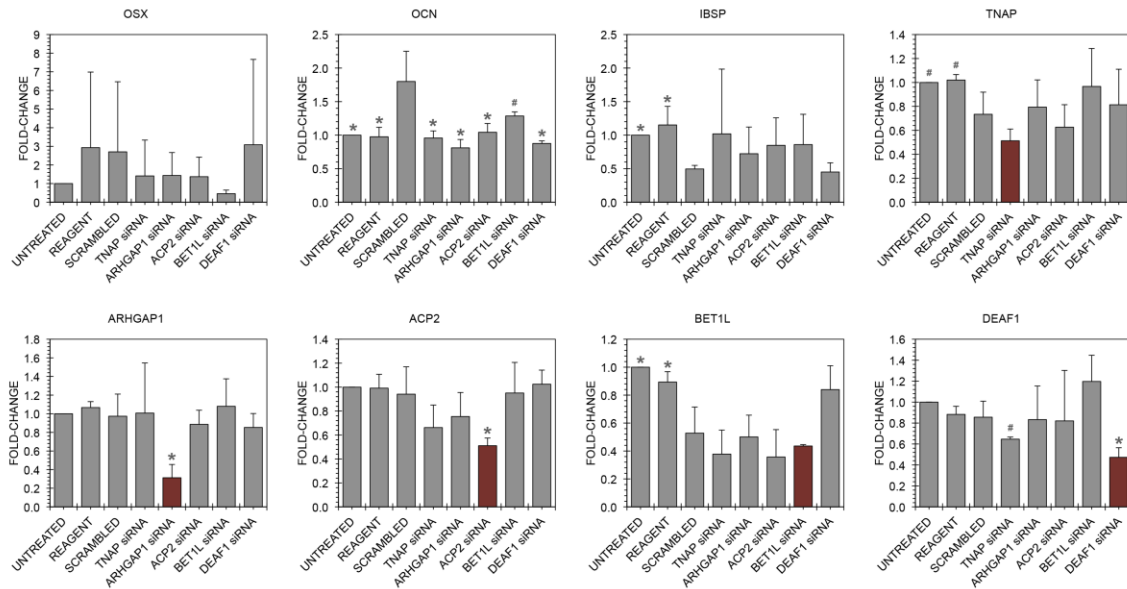
Figure 3: *Real-time PCR of osteoblast differentiation genes and GWAS/TAD hits in hFOB cells.*
siRNA was used to knock down expression of *TNAP* (positive control), *ARHGAP1*, *ACP2*,
BET1L and *DEAF1*. Relative expression of the osteoblast marker genes *OSX*, *OCN* and *IBSP*
suggest that GWAS/TAD hits 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* ≤
0.05 and #*p* ≤ 0.10 using a two-tailed Student's *t*-test.

210
211
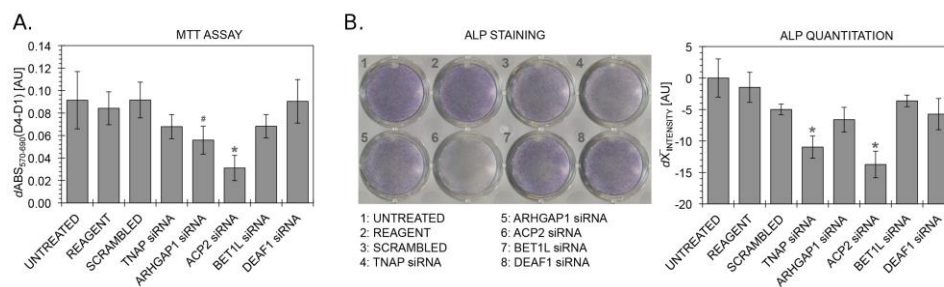
212
Figure 4. *Validating two 'TAD Pathway' predictions for Bone Mineral Density GWAS hits on
hFOB cells.* siRNA was used to knock down expression of *TNAP*, *ARHGAP1*, *ACP2*, *BET1L* and
DEAF1. (**A**) Knockdown of *ACP2* decreases cellular metabolic activity, demonstrated using an
MTT assay. (**B**) 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* ≤
0.05 and #*p* ≤ 0.10 using a two-tailed Student's *t*-test.

220

10

221 *Influence of TAD_Pathways Gene Predictions on Alkaline Phosphatase Activity*

222   Alkaline phosphatase (ALP) is highly expressed in osteoblasts; disruption of proliferation or

223 osteoblast differentiation would result in downregulation of ALP. Treatment with siRNA

224 resulted in changes in ALP staining that we analyzed further by quantitation. *TNAP* siRNA

225 significantly reduced ALP by 5.98±1.77 versus the scrambled siRNA control ($p = 0.006$). *ACP2*

226 siRNA also significantly reduced ALP intensity by 8.74±2.11 versus the scrambled siRNA

227 control ($p = 0.003$). The scrambled siRNA group stained less intensely than untreated or

228 transfection reagent control wells, but this did not reach statistical significance ($0.05 < p < 0.10$)

229 (**Figure 4B**).

230

231 **Discussion:**

232   We observed a nonrandom enrichment of SNPs in the center of TADs that was consistent

233 across different cell types, but was in the opposite direction of the gene and repeat elements

234 distributions. It is possible that the gene distribution is driving this phenomenon, since coding

235 regions are under higher evolutionary constraint and are thus more averse to SNPs [39].

236 Nevertheless, GWAS SNPs also appeared to be distributed closely to boundary regions in hESC

237 cells. This may support GWAS causally implicating nearest genes more frequently since genes

238 are also distributed near boundaries. The observation may also suggest that polymorphism in

239 regions near TAD boundaries are more important drivers of disease risk associations than

240 polymorphism in the center of TADs. However, we do not observe this pattern in IMR90 cells.

241 The SNP distribution was also opposite of the SINE/Alu repeat distribution. Given that Alu

242 elements tend to insert into GC-rich regions [40], we tracked GC content across TADs and

243 observed only a slight increase in GC content near TAD boundaries. There was also a slightly

11

244   inverse distribution of Alu evolutionary divergence [40]. Our results suggest that the Alu

245   distribution is primarily driven by intronic clustering [41] rather than GC-biased insertion or

246   evolutionary divergence. Recently, retrotransposons have been shown to act as genomic

247   insulators [42], while Alu repeats have been shown to be correlated with functional elements

248   [43]. However, the relationships between polymorphic sites, repeat elements, and genes across

249   TADs and higher level genome organization have yet to be explored in detail and warrants

250   further investigation.

251   TAD boundaries offer a unique computational opportunity to use biologically informed

252   windows to predefine areas of the genome that are more likely to interact with themselves. We

253   showed, as a proof of concept, that TADs can reveal functional GWAS variant to gene

254   relationships using BMD. Several of the TAD_Pathways implicated genes, including *LRP5* and

255   other Wnt signaling genes, are *bona fide* BMD genes already identified by nearest gene GWAS,

256   eQTL analyses and human clinical syndromes [44,45], thus providing positive controls for our

257   approach. However, several BMD GWAS signals do not have obvious nearest gene associations,

258   which allowed us to validate our approach with two candidate causal, non-nearest gene

259   predictions: *ACP2* and *DEAF1*. Both genes also did not have eQTL associations, but this is

260   likely a result of the eQTL browser lacking bone tissue.

261   To assess the validity of our predictions, we experimentally knocked down *ACP2* and

262   *DEAF1* in hFOB cells. siRNA for *ACP2* and *DEAF1* did not significantly alter expression of the

263   osteoblast marker genes *OSX*, *IBSP* or *TNAP*. *OCN* was downregulated in each of the

264   experimental groups relative to the scrambled siRNA control, but comparison with the reagent

265   control indicated no significant difference in any group, suggesting an off-target effect on *OCN*

266   in the scrambled siRNA group. Because osteoblast differentiation genes were not downregulated

267   following knockdown of the genes of interest, we concluded that these genes do not directly

268   regulate the transcriptional processes of osteoblast differentiation *in vitro*. The decrease in

269   *DEAF1* expression following *TNAP* siRNA treatment, though not statistically significant,

270   suggests that *DEAF1* may function downstream of *TNAP*.

271   There was a pronounced and statistically significant reduction in metabolic activity in hFOB

272   cells treated with *ACP2* siRNA. This result carried through to the ALP assay, in which staining

273   intensity and ALP+ area fraction were dramatically reduced in only the *TNAP* siRNA and *ACP2*

274   siRNA groups. The combination of these results with the gene expression data suggests that

275   *ACP2* regulates early osteoblast proliferation/viability, but does not directly regulate osteoblast

276   differentiation.

277   We provide evidence that our approach can steer researchers from GWAS signals toward

278   genes relevant to the pathogenesis of the given trait. Furthermore, because our method treats all

279   genes in implicated TADs equally, functional classification extends to the identification of single

280   variant pleiotropic events; as was the case with an intronic *FTO* variant impacting both *IRX3* and

281   *IRX5* [8].

282   Despite the advantages presented by TAD_Pathways, the method has a number of

283   limitations. Currently, our method will not overcome the possibility of a gene being

284   inappropriately included in a pathway that it does not actually contribute to, plus all other

285   propagated errors related to pathway curation and analyses [46]. Network based methods built on

286   gene-gene interaction data also suffer from similar biases [47], but potentially to a lesser extent

287   than curated pathways. We include both curated and computationally predicted GO annotations

288   to ameliorate this bias. The computational predictions provide additional support that these genes

289   may be important disease associated genes that we would have missed using only experimentally

290   validated pathway genes. We are also unable to implicate a gene to a trait if it is not assigned to a

291   curated or predicted pathway, or if it does not fall within a TAD corresponding to a GWAS

292   signal. It is also likely that our approach will not work well with every GWAS. Indeed we are

293   implicating causality to given genes - we are not making a direct connection between the gene

294   and the given variant. Furthermore, our method does not include the possibility of finding genes

295   associated with a disease that is impacted by alternative looping, which has been observed to

296   occur in cancer [48,49] and sickle cell anemia [50]. As research on 3D genome organization

297   increases, it is likely that more diseases will include chromosome looping deficiencies as part of

298   their etiology. Additionally, we used TAD boundaries defined by Dixon *et al.* 2012. A more

299   recent Hi-C analysis at increased resolution substantially reduced the estimated average size of

300   TADs [51]. Nevertheless, there remains disagreement about how TADs are defined [52]. Despite

301   our method using larger TAD boundaries, thus promoting the inclusion of more presumably false

302   positive genes, we retain the ability to identify biologically logical pathways. The larger

303   boundaries permit us to screen a larger number of candidate genes but makes the method

304   analytically conservative by increasing the pathway overrepresentation signal required to surpass

305   the adjusted significance threshold.

306   The validation screen is also limited: it was performed in a simplified *in vitro* cell culture

307   system lacking organismal complexity, and the cell line selected is largely tetraploid which may

308   partially compensate for gene knockdown. This is particularly true for the lack of reduction in

309   *TNAP* gene expression in the *TNAP* siRNA group, in light of the historical selection of the hFOB

310   cell line based on robust ALP staining in culture [53]. As well, while the TAD approach

311   identifies potentially several GWAS associated genes, herein we only examined two genes per

312   TAD – one immediately adjacent to the GWAS SNP and another that we postulated could play a

313   role in the skeleton. Further work would need to systematically examine the relative importance

314   of each gene in a TAD.

315   Other recent mechanisms and algorithms used to assign causality from association signals or

316   enhancers to genes typically leverage multiple data types including expression [54] or epigenetic

317   features [24]. For example, TargetFinder uses several high throughput genomic marks to identify

318   features predictive of a chromosome physically looping together enhancers and promoters [24].

319   Looping occurs at sub-TAD level resolutions [55] and sub-TADs are variable across cell types.

320   Therefore, in order for a chromosome looping signature to generalize to GWAS signals, a user

321   must assay tissue-specific and high resolution Hi-C to identify more specific interactions.

322   Alternatively, one could also query variants that affect gene expression in high-throughput and

323   systematically match signals to gene expression [56]. A major limitation to these approaches is

324   that several diseases do not yet have a known tissue source or involve multiple tissues. This is

325   particularly true for osteoporosis whereby multiple cell types, as well as systemic factors,

326   influence bone mass [57]. Therefore, identifying these specific signals may require the

327   procedures to be repeated across competing tissue types. In contrast, a TAD_Pathways analysis

328   is computationally cheap and uses publicly available TAD boundaries and GO terms as a guide

329   for assigning genes to GWAS signals. The method is effective in a wide variety of settings and

330   across tissues because TADs are consistent across cell types [25]. In summary, TAD_Pathways

331   can be used to guide researchers toward the most likely causal gene implicated by a GWAS

332   signal. We have also identified *ACP2* as a gene involved in BMD determination, which warrants

333   further investigation.

334

335   **Conclusions:**

336   TADs offer a novel tool in the investigation of genome function. We present an approach,

337   called TAD_Pathways, to leverage 3D genomics to prioritize and predict causal genes implicated

338   by GWAS signal. At the foundation of our method is the principle that genomic regions within

339   the same TAD more often interact with each other, and therefore, provide the genomic

340   scaffolding that can impact gene function and gene regulation within each TAD. We applied our

341   method to established BMD GWAS signals. By selecting two GWAS signals and two classes of

342   genes for each signal (nearest gene and predicted gene by TAD_Pathways), we demonstrated

343   that our approach can causally implicate genes kilobases away from their associated GWAS

344   signal. We validated *ACP2* (TAD prediction), but not *ARHGAP1* (nearest gene), and show that

345   *ACP2* influences the proliferation and differentiation of osteoblast cells. We were unable to

346   validate either *DEAF1* or *BET1L* and conclude that neither impacts osteoblast gene expression

347   nor metabolic activity. Whether these genes influence other aspects of skeletal biology cannot be

348   determined within the scope of the current study. Future studies focused on BMD GWAS would

349   explore both osteoblast and osteoclast associated changes. In conclusion, as more information

350   and data is collected regarding 3D genome principles, we propose that algorithms that leverage

351   dynamic 3D structure rather than static linear organization will more accurately predict and

352   discover the basic genomic biology of diseases.

353

354   **Methods:**

355   *Data Integration*

356   We used previously identified TAD boundaries for hESC, IMR90, mESC, and mcortex cells

357   for all TAD based analyses [22,58]. To describe the genomic content of TADs, we extracted

358   common SNPs (major allele frequency $\leq$ 0.95) from the 1000 Genomes Phase III data (2 May

16

359   2013 release) [59] and downloaded hg19 Gencode genes [60] and hg19 RepeatMasker repeat

360   elements [61]. We downloaded hg19 FASTA files for all chromosomes as provided by the

361   Genome Reference Consortium [62]. Furthermore, we downloaded the NHGRI-EBI GWAS

362   catalog on 25 February 2016, which holds the significant findings of several GWAS' for over

363   300 traits [1]. Since the GWAS catalog reports hg38 coordinates, we used the hg38 to hg19

364   UCSC chain file [63] and PyLiftover [64] to convert genome build coordinates to hg19. We

365   assessed relevant expression quantitative trait loci (eQTLs) using all tissues in the NCBI GTEx

366   eQTL Browser [65].

367

368   *TAD_Pathways*

369   Our TAD_Pathways method is a light-weight approach that uses TAD boundary regions,

370   rather than distance explicitly, to identify putative causal genes. We first build a comprehensive

371   TAD based gene list that consists of all genes that fall inside TADs that are implicated with a

372   GWAS signal (see **Figure 2**). This gene list assumes that all genes within each signal TAD have

373   an equal likelihood of functional impact on the trait or disease of interest. We then input the

374   TAD based gene list into a WebGestalt overrepresentation analysis [66]. WebGestalt is a

375   webapp that facilitates a pathway analysis interface allowing for quick and custom gene set

376   based analyses. We perform a pathway overrepresentation test for the input TAD based genes

377   against GO biological process, molecular function, and cellular component terms with a

378   background of the human genome. Specifically, this tests if the input gene set is associated with

379   any particular GO term at a higher probability than by chance compared to background genes.

380   We include both experimentally validated and computationally inferred genes in each GO term,

381   which allows the method to discover associations for genes that lack literature support. We

382   consider genes that are annotated to the most significantly enriched GO term to be the associated

383   set [65].

384

385   *Cell culture and siRNA transfection*

386   A human fetal osteoblast cell line (ATCC hFOB 1.19 CRL-11372) was obtained and

387   subcultured twice at 34°C, 5% CO_2 and 95% relative humidity in 1:1 DMEM/F12 with 2.5mM

388   L-glutamine without phenol red (Gibco 21041025) supplemented with 10% FBS (Atlas USDA

389   F0500D) and 0.3mg/mL G418 sulfate (Gibco 10131035). All experiments were conducted in

390   three temporally separated independent technical replicates from cryopreserved P2 aliquots of

391   these cells. 48 hours prior to transfection, media was switched to a G418-free formulation.

392   Transfections were conducted in single-cell suspension using a commercial siRNA reagent

393   system (Santa Cruz sc-45064) according to the manufacturer's instructions, with 6μL of siRNA

394   duplex and 4μL transfection reagent in 750μL of transfection media per 100,000 cells. Following

395   trypsinization, cells were counted and divided into one of 8 experimental groups: 1) untreated

396   control, 2) siRNA-negative transfection reagent control, 3) scrambled control siRNA (sc-37007),

397   4) *TNAP* siRNA (sc-38921), 5) *ARHGAP1* siRNA (sc-96477), 6) *ACP2* siRNA (sc-96327), 7)

398   *BET1L* siRNA (sc-97007) or 8) Suppressin (*DEAF1*) siRNA (sc-76613). Samples for RNA

399   isolation were generated by plating 200,000 cells per well in tissue culture treated 6-well plates

400   (Falcon 353046). Samples for MTT assay and alkaline phosphatase (ALP) staining were

401   generated by plating 50,000 cells per well in tissue culture treated 24-well plates (Falcon

402   353047). Cells were then switched to a 37°C incubator and transfected for 6 hours, at which

403   point transfection cocktails were diluted with 2x hFOB media concentrate. Media was

404   completely changed to 1x standard hFOB media 16 hours later.

405

Quantitative PCR

406

407   RNA was harvested at Day 4 using TRIzol (Ambion) followed by acid guanidinium

408   thiocyanate-phenol-chloroform extraction and RNeasy (Qiagen) spin-column purification with

409   DNase (Qiagen) then reverse-transcribed using a high-capacity RNA-to-cDNA kit (Applied

410   Biosystems). Duplicate qPCR reactions were conducted on 20ng of whole-RNA template using

411   SYBR Select master mix (Applied Biosystems) in an Applied Biosystems 7500 Fast Real-Time

412   PCR System. Primers for *GAPDH*, *OSX*, *OCN* and *IBSP* were adopted from the literature and

413   synthesized by R&D Systems. New primer sets spanning exon-exon junctions were designed for

414   *ARHGAP1*, *ACP2*, *BET1L* and *DEAF1* using NCBI Primer Blast, verified by melt curve analysis

415   and agarose gel electrophoresis. Sequences follow: *ARHGAP1* F-

416   GCGGAAATGGTTGGGGATAG R-CCTTAAGAGAAACCGCGCTC (127bp), *ACP2* F-

417   AGCGGGTTCCAGCTTGTTT R-TGGCGGTACAGCAAGGTAAC (165bp), *BET1L* F-

418   GGATGGCATGGACTCGGATT R-TCCTCTGGAGCCCAAAACAC (254bp), *DEAF1* F-

419   GGAAGGAGCAGTCCTGCGTT R-TCACCTTCTCCATCACGCTTT (195bp). Results were

420   analyzed using the 2^{-ddCt} method using GAPDH as a housekeeping gene and reported as (mean ±

421   standard deviation) fold-change versus untreated controls. Statistical significance was

422   determined using 2-way homoscedastic Student's t-tests versus the scrambled siRNA control,

423   annotated using $*p \leq 0.05$ and $\#p \leq 0.10$. The acronym N.S. stands for "not significant".

424

MTT metabolic assay

425

426   Cellular metabolism/proliferation was assessed using a commercial cell growth

427   determination kit (Sigma CGD1). At 1 or 4 days post-transfection, media in assigned 24-well

428   plates was switched to 450μL hFOB media plus 50μL MTT solution. Cells were incubated at

429   37°C for 3.5 hours, after which the media was aspirated and the resulting formazan crystals were

430   solubilized in MTT solvent. Plates were shaken and read in a BioTek Synergy H1 microplate

431   reader. Results are reported as the difference in mean absorbance at 570nm-690nm from Day 1

432   to Day 4. Error bars represent root-mean-square standard deviation from the measurements at

433   both days. Statistical significance was determined using 2-way homoscedastic Student's t-tests

434   versus the scrambled siRNA control, annotated using $*p \leq 0.05$ and $\#p \leq 0.10$.

435

436   *Alkaline phosphatase (ALP) staining*

437   Plates were stained at 4 days post-transfection using a commercial ALP kit (Sigma 86C).

438   Dried plates were then imaged at 600dpi using an Epson V370 photo scanner and staining was

439   quantified using mean whole-well intensity measurements in ImageJ using raw output files. ALP

440   area fraction was calculated using color thresholding. Values are reported as mean ± standard

441   deviation. Statistical significance was determined using 2-way homoscedastic Student's t-tests

442   versus the scrambled siRNA control, annotated using $*p \leq 0.05$ and $\#p \leq 0.10$. Images of plates

443   presented as figures were edited using a warming filter and for brightness/contrast using Adobe

444   Photoshop CS6.

445

446   *Computational Reproducibility*

447   We provide all of our source code under a permissive open source license and encourage

448   others to modify and build upon our work [67]. Additionally, we provide an accompanying

449   docker image [68] to replicate our computational environment

450   (https://hub.docker.com/r/gregway/tad_pathways/) [69].

451

Declarations:

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

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).

Competing interests

The authors declare no competing interests.

Funding

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.

Authors' contributions

GPW wrote the software, analyzed the data, developed the method and wrote the manuscript; DWY performed the experimental validation and wrote the manuscript; KDH performed the

474   experimental validation and wrote the manuscript; CSG analyzed the data, developed the

475   method and wrote the manuscript; SFAG analyzed the data, developed the method and wrote

476   the manuscript.

477 *Acknowledgements*

478   Hannah E. Sexton and Troy L. Mitchell assisted in optimizing siRNA transfection

479   conditions. Daniel Himmelstein and Amy Campbell performed analytical code review.
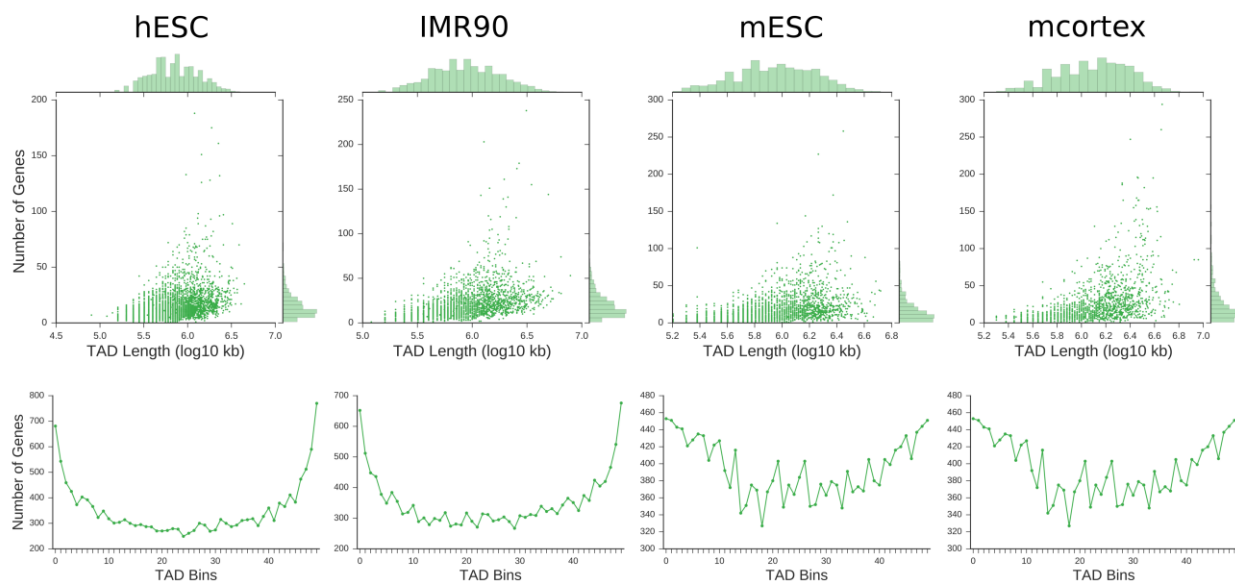
480 *Authors' information (optional)*

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497   **Supplementary Figures:**

498

499   **Supplementary Figure S1:** *Distributions of genes across topologically associating domains of*
500   *four cell types.* Bars on the top and right side of the plots represent histograms of each respective
501   metric. We independently discretized TADs into 50 bins where the distribution of elements is
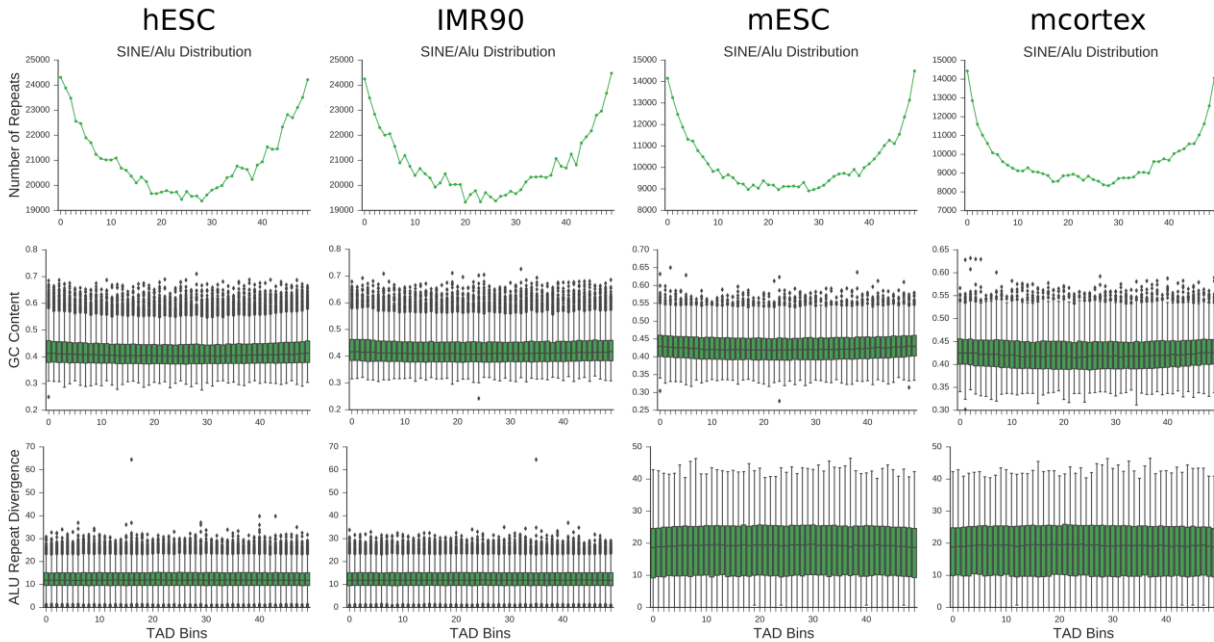502   linear from 5' to 3' (bin 0 is the 5' most end of all TADs).

503
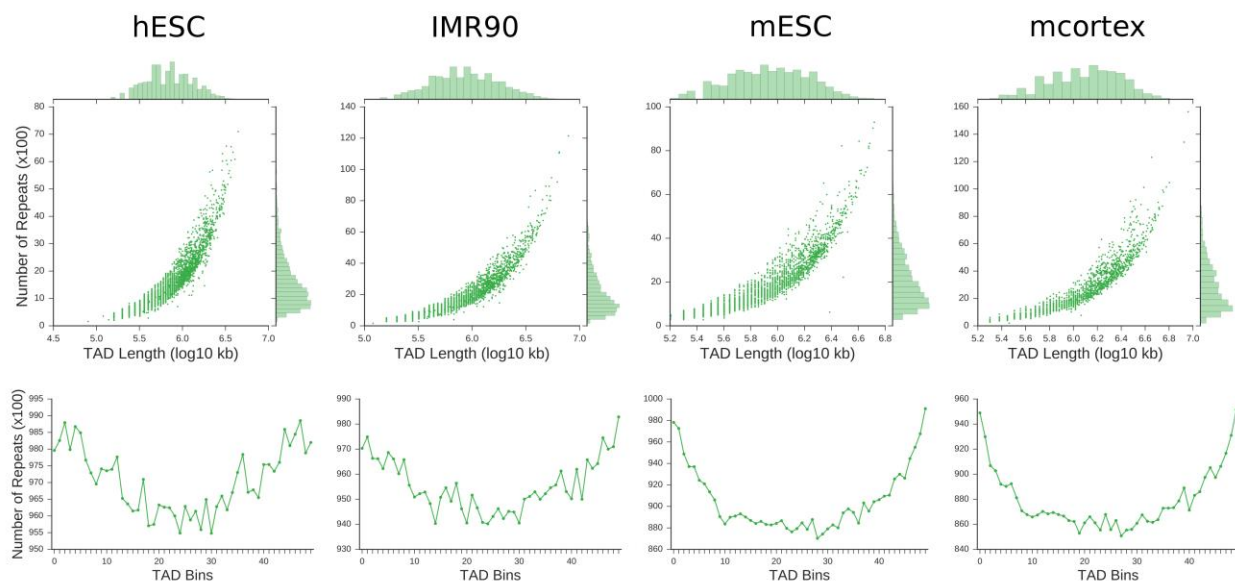
504

505   **Supplementary Figure S2:** *Distributions of all repeat elements across topologically associating*
506   *domains of four cell types.* Bars on the top and right side of the plots represent histograms of
507   each respective metric. We independently discretized TADs into 50 bins where the distribution
508   of elements is linear from 5' to 3' (bin 0 is the 5' most end of all TADs).

509

510

Supplementary Figure S3: *Distributions of SINE/Alu elements and potential confounding factors across topologically associating domains of four cell types.* We independently discretized TADs into 50 bins where the distribution of elements is linear from 5' to 3' (bin 0 is the 5' most end of all TADs). The distribution of Alu elements is strongly clustered toward TAD boundaries but neither the percentage of GC content across TADs nor the evolutionary divergence of Alu elements can explain the Alu distribution.

517

518
519
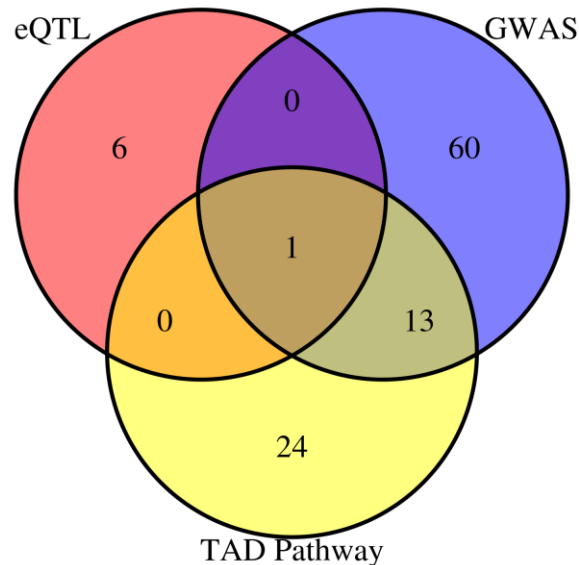
Supplementary Figure S4: *Bone Mineral Density gene counts associated with various classes of evidence.* The gene *LRP5* is implicated by all evidence codes. About 1/3 of all TAD_Pathways implicated genes are the nearest gene implicated by the GWAS signal.

24

523   **Tables:**

524   **Supplementary Table S1:** *Chi square testing the enrichment of genes at TAD boundaries and*

525   *enrichment of SNPs towards the right half of TADs.*

	Enrichment of SNPs in the Right of TADs			
	Left	Right	Chi-square	P
hESC	3268781	3296450	116.6	3.5E-27
IMR90	3211140	3210251	0.12	0.73
mESC*	25161934	25129279	21.2	4.1E-06
mcortex	24616532	24641931	13.1	3.0E-4

526   * Note- mouse embryonic stem cells have an opposite enrichment. There are more SNPs in the

527   left of TADs

528

529   We provide **Supplementary Tables S2** and **S3** as attached .xls files.

530

531   **Abbreviations:**

532   TAD – Topologically associating domain

533   GWAS – Genome wide association study

534   SNP – Single nucleotide polymorphism

535   BMD – Bone mineral density

536   hESC – Human embryonic stem cells

537   mESC – Mouse embryonic stem cells

538   IMR90 – Human fibroblast cells

539   mcortex – Mouse cortex cells

540   siRNA – Small interfering RNA

541   eQTL – Expression quantitative trail loci

542   hFOB – Human fetal osteoblast

25

543   TNAP - tissue-nonspecific alkaline phosphatase

544   OCN – osteocalcin

545   IBSP – bone sialoprotein

546   ALP – Alkaline phosphatase

547   GO – Gene ontology

548   eQTL – Expression quantitative trait loci

549

550 **References:**

551 1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS
552 Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–6.

553 2. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum. Mol. Genet.
554 2015;24:R102–10.

555 3. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human
556 disease. Nat. Biotechnol. 2012;30:1095–106.

557 4. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR,
558 Chakravarti A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature.
559 2012;491:56–65.

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

562 6. Tung YCL, Yeo GSH, O'Rahilly S, Coll AP. Obesity and FTO: Changing Focus at a Complex
563 Locus. Cell Metab. 2014;20:710–8.

564 7. Ragvin A, Moro E, Fredman D, Navratilova P, Drivenes O, Engstrom PG, et al. Long-range
565 gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and
566 IRX3. Proc. Natl. Acad. Sci. 2010;107:775–80.

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

569 9. Smemo S, Tena JJ, Kim K-H, Gamazon ER, Sakabe NJ, Gómez-Marín C, et al. Obesity-
570 associated variants within FTO form long-range functional connections with IRX3. Nature.
571 2014;507:371–5.

572   10. Grant SFA, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, et al.
573   Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat.
574   Genet. 2006;38:320–3.

575   11. Xia Q, Chesi A, Manduchi E, Johnston BT, Lu S, Leonard ME, et al. The type 2 diabetes
576   presumed causal variant within TCF7L2 resides in an element that controls the expression of
577   ACSL5. Diabetologia. 2016;59:2360–8.

578   12. Herman MA, Rosen ED. Making Biological Sense of GWAS Data: Lessons from the FTO
579   Locus. Cell Metab. 2015;22:538–9.

580   13. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science.
581   2002;295:1306–11.

582   14. Dekker J. Gene Regulation in the Third Dimension. Science. 2008;319:1793–4.

583   15. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al.
584   Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human
585   Genome. Science. 2009;326:289–93.

586   16. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, et al. A three-dimensional
587   model of the yeast genome. Nature. 2010;465:363–7.

588   17. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, et al.
589   Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage
590   Commitment. Cell. 2013;153:1281–95.

591   18. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-
592   Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell.
593   2012;148:458–72.

594   19. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and
595   topological characteristics of mammalian regulatory domains. Genome Res. 2014;24:390–400.

596   20. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, et al. Topologically associating
597   domains are stable units of replication-timing regulation. Nature. 2014;515:402–5.

598   21. Gurudatta B, Yang J, Van Bortle K, Donlin-Asp P, Corces V. Dynamic changes in the
599   genomic localization of DNA replication-related element binding factor during the cell cycle.
600   Cell Cycle. 2013;12:1605–15.

601   22. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in
602   mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.

603   23. Nora EP, Dekker J, Heard E. Segmental folding of chromosomes: A basis for structural and
604   regulatory chromosomal neighborhoods?: Prospects & Overviews. BioEssays.
605   2013;35:818–28.

606 24. Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex
607 genomic signatures on looping chromatin. Nat. Genet. 2016;48:488–96.

608 25. Smith EM, Lajoie BR, Jain G, Dekker J. Invariant TAD Boundaries Constrain Cell-Type-
609 Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus.
610 Am. J. Hum. Genet. 2016;98:185–201.

611 26. Richards JB, Rivadeneira F, Inouye M, Pastinen TM, Soranzo N, Wilson SG, et al. Bone
612 mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study.
613 Lancet Lond. Engl. 2008;371:1505–12.

614 27. Rivadeneira F, Styrkársdottir U, Estrada K, Halldórsson BV, Hsu Y-H, Richards JB, et al.
615 Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide
616 association studies. Nat. Genet. 2009;41:1199–206.

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

620 29. Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, Jonasdottir A, et
621 al. Nonsense mutation in the LGR4 gene is associated with several human diseases and other
622 traits. Nature. 2013;497:517–20.

623 30. Hendrickx G, Boudin E, Van Hul W. A look behind the scenes: the risk and pathogenesis of
624 primary osteoporosis. Nat. Rev. Rheumatol. 2015;11:462–74.

625 31. Choo SY. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical
626 Implications. Yonsei Med. J. 2007;48:11.

627 32. Håvik B, Le Hellard S, Rietschel M, Lybæk H, Djurovic S, Mattheisen M, et al. The
628 Complement Control-Related Genes CSMD1 and CSMD2 Associate to Schizophrenia. Biol.
629 Psychiatry. 2011;70:35–42.

630 33. Kwon E, Wang W, Tsai L-H. Validation of schizophrenia-associated genes CSMD1,
631 C10orf26, CACNA1C and TCF4 as miR-137 targets. Mol. Psychiatry. 2013;18:11–2.

632 34. Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide
633 Association Studies. Am. J. Hum. Genet. 2007;81:1278–83.

634 35. Hao K, Di X, Cawley S. LdCompare: rapid computation of single- and multiple-marker r2
635 and genetic coverage. Bioinformatics. 2007;23:252–4.

636 36. Taşan M, Musso G, Hao T, Vidal M, MacRae CA, Roth FP. Selecting causal genes from
637 genome-wide association studies via functionally coherent subnetworks. Nat. Methods.
638 2014;12:154–9.

639 37. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology:
640 tool for the unification of biology. Nat. Genet. 2000;25:25–9.

641 38. Petersen A, Alvarez C, DeClaire S, Tintle NL. Assessing Methods for Assigning SNPs to
642 Genes in Gene-Based Tests of Association Using Common Variants. Chen L, editor. PLoS ONE.
643 2013;8:e62161.

644 39. Mu XJ, Lu ZJ, Kong Y, Lam HYK, Gerstein MB. Analysis of genomic variation in non-
645 coding elements using population-scale sequencing data from the 1000 Genomes Project.
646 Nucleic Acids Res. 2011;39:7058–76.

647 40. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Duplication, coclustering, and
648 selection of human Alu retrotransposons. Proc. Natl. Acad. Sci. 2004;101:1268–72.

649 41. Deininger P. Alu elements: know the SINEs. Genome Biol. 2011;12:236.

650 42. Wang J, Vicente-García C, Seruggia D, Moltó E, Fernandez-Miñán A, Neto A, et al. MIR
651 retrotransposon sequences provide insulators to the human genome. Proc. Natl. Acad. Sci.
652 2015;112:E4428–37.

653 43. Gu Z, Jin K, Crabbe MJC, Zhang Y, Liu X, Huang Y, et al. Enrichment analysis of Alu
654 elements with different spatial chromatin proximity in the human genome. Protein Cell.
655 2016;7:250–66.

656 44. Baron R, Kneissel M. WNT signaling in bone homeostasis and disease: from human
657 mutations to treatments. Nat. Med. 2013;19:179–92.

658 45. Krishnan V, Bryant HU, Macdougald OA. Regulation of bone mass by Wnt signaling. J.
659 Clin. Invest. 2006;116:1202–9.

660 46. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and
661 Outstanding Challenges. Ouzounis CA, editor. PLoS Comput. Biol. 2012;8:e1002375.

662 47. Reguly T, Breitkreutz A, Boucher L, Breitkreutz B-J, Hon GC, Myers CL, et al.
663 Comprehensive curation and analysis of global interaction networks in Saccharomyces
664 cerevisiae. J. Biol. 2006;5:11.

665 48. Corces MR, Corces VG. The three-dimensional cancer genome. Curr. Opin. Genet. Dev.
666 2016;36:1–7.

667 49. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO,
668 et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature.
669 2015;529:110–4.

670 50. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, et al. Controlling long-range
671 genomic interactions at a native locus by targeted tethering of a looping factor. Cell.
672 2012;149:1233–44.

673 51. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D
674 Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.
675 Cell. 2014;159:1665–80.

676 52. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat. Rev. Genet.
677 2016;17:661–78.

678 53. Harris SA, Enger RJ, Riggs BL, Spelsberg TC. Development and characterization of a
679 conditionally immortalized human fetal osteoblastic cell line. J. Bone Miner. Res. Off. J. Am.
680 Soc. Bone Miner. Res. 1995;10:178–86.

681 54. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary
682 data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet.
683 2016;48:481–7.

684 55. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, et al. Control of Cell Identity
685 Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell. 2014;159:374–
686 87.

687 56. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct Identification of
688 Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell.
689 2016;165:1519–29.

690 57. Lupsa BC, Insogna K. Bone Health and Osteoporosis. Endocrinol. Metab. Clin. North Am.
691 2015;44:517–30.

692 58. Ho JWK, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, et al. Comparative analysis of
693 metazoan chromatin organization. Nature. 2014;512:449–52.

694 59. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang
695 HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

696 60. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al.
697 GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res.
698 2012;22:1760–74.

699 61. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. Available from:
700 http://www.repeatmasker.org

701 62. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing
702 Reference Genome Assemblies. PLoS Biol. 2011;9:e1001091.

703 63. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human
704 Genome Browser at UCSC. Genome Res. 2002;12:996–1006.

705 64. Tretyakov K. pyliftover [Internet]. 2014. Available from:
706 https://github.com/konstantint/pyliftover

707 65. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet.
708 2013;45:580–5.

709 66. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt):
710 update 2013. Nucleic Acids Res. 2013;41:W77–83.

711 67. Gregory Way, Casey Green. Greenelab/Tad_Pathways: Pre-Release. 2016 [cited 2016 Nov
712 14]; Available from: https://doi.org/10.5281/zenodo.163950

713 68. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Oper. Syst.
714 Rev. 2015;49:71–9.

715 69. Gregory Way, Struan Grant, Casey Greene. TAD Pathways Archived Docker Image. 2016
716 [cited 2016 Nov 14]; Available from: https://doi.org/10.5281/zenodo.166556

717