Discovery of gene regulatory elements through a new bioinformatics analysis of haploid genetic screens

Bhaven B. Patel[1,2¶], Andres M. Lebensohn[1,2¶*], Jan E. Carette[3], Julia Salzman[1,4*], and Rajat Rohatgi[1,2*]

[1] Department of Biochemistry, Stanford University School of Medicine, Stanford, California, United States of America

[2] Department of Medicine, Stanford University School of Medicine, Stanford, California, United States of America

[3] Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, California, United States of America

[4] Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California, United States of America

* Corresponding authors
E-mail: andreslebensohnphd@gmail.com (AML), julia.salzman@stanford.edu (JS) and rrohatgi@stanford.edu (RR)

¶These authors contributed equally to this work.

Abstract

The systematic identification of regulatory elements that control gene expression remains a challenge. Genetic screens that use untargeted mutagenesis have the potential to identify protein-coding genes, non-coding RNAs and regulatory elements, but their analysis has mainly focused on identifying the former two. To identify regulatory elements, we conducted a new bioinformatics analysis of insertional mutagenesis screens interrogating WNT signaling in haploid human cells. We searched for specific patterns of retroviral gene trap integrations (used as mutagens in haploid screens) in short genomic intervals overlapping with introns and regions upstream of genes. We uncovered atypical patterns of gene trap insertions that were not predicted to disrupt coding sequences, but caused changes in the expression of two key regulators of WNT signaling, suggesting the presence of cis-regulatory elements. Our methodology extends the scope of haploid genetic screens by enabling the identification of regulatory elements that control gene expression.

Introduction

An outstanding challenge in genomics is the identification of functional regulatory elements that control spatial and temporal expression of protein-coding genes and non-coding RNAs. The Encyclopedia of DNA Elements (ENCODE) project has the ambitious goal of generating a candidate list of all functional elements in the human genome using sequence features, such as conservation, and biochemical features, such as chromatin accessibility and chromatin modifications [1]. Functional approaches to identify regulatory elements have thus far focused on specific regions of the genome and include massively parallel reporter assays or dense clustered regularly-interspaced short palindromic repeats (CRISPR)-mediated mutagenesis of <1 megabase pair segments around a locus of interest (reviewed in [2]).

In work published recently [3], we conducted a comprehensive set of forward genetic screens in haploid human cells to uncover genes required for signaling through the WNT pathway, which plays central roles in development, stem cell function, and cancer. The power of these screens, which used a quantitative transcriptional reporter as the basis for phenotypic selection, was highlighted by the identification of genes encoding both known and novel components that function at most levels of the WNT pathway, from the cell surface to the nucleus. Our previous analysis focused primarily on annotated protein-coding genes and non-coding RNAs. Since the mutant cell libraries used in these screens were generated through untargeted mutagenesis of the genome with a gene trap (GT)-bearing retrovirus, we wondered whether we could use the datasets generated by these screens to uncover gene regulatory mechanisms that modulate the WNT signaling pathway. While retroviral insertions can happen throughout the genome, they are most predominant around transcriptional start sites (TSS),

63 promoters, and enhancers [4]; hence we focused our analysis of retroviral GT insertions on non-
64 coding regions in genes and immediately upstream of them.
65 Here we present a new bioinformatics pipeline designed to uncover gene regulatory
66 elements and provide evidence for regulatory regions in the first intron of the gene encoding the
67 transcription factor AP4 (TFAP4), a positive regulator of WNT signaling [3], and in the genomic
68 region upstream of the promoter for the gene encoding the WNT co-receptor LRP6.

Materials and Methods

Reagents

The reagents used in this manuscript are described in the Materials and methods of [3] and below.

Antibodies

For immunoblotting:

Primary antibodies: rabbit anti-AP4 (TFAP4) serum (1:2000, a gift from Takeshi Egawa [5]); rabbit anti-LRP6 (C5C7) (1:500, Cell Signaling Technologies Cat. # 2560); mouse anti-ACTIN (clone C4) (1:500, EMD Millipore Cat. # MAB1501).

Secondary antibodies: peroxidase AffiniPure goat anti-rabbit IgG (H+L) (1:7500, Jackson ImmunoResearch Laboratories Cat. # 111-035-003); IRDye 800CW donkey anti-rabbit IgG (H+L) (1:10,000, Li-Cor Cat. # 925-32213); IRDye 800CW donkey anti-mouse IgG (H+L) (1:10,000, Li-Cor Cat. # 926-32212).

Primary and secondary antibodies used for detection with the Li-Cor Odyssey imaging system were diluted in a 1 to 1 mixture of Odyssey Blocking Buffer (Li-Cor Cat. # 927–40000) and TBST (Tris buffered saline (TBS) + 0.1% Tween-20), and those used for detection by chemiluminescence were diluted in TBST + 5% skim milk. Primary antibody incubations were

90 done overnight at 4°C, and secondary antibody incubations were done for 1 hr at room
91 temperature (RT).
92
93 **For immunostaining:**
94 Primary antibodies: mouse anti-LRP6 (clone A59) (5µg/mL, Millipore Cat. # MABS274).
95
96 Secondary antibodies: donkey anti-mouse IgG (H+L) Alexa Fluor 647 conjugate (1:200, Thermo
97 784 Fisher Scientific Cat. # A-31571).
98

99 Cell lines and growth conditions

100 WT HAP1-7TGP cells and genetically modified clonal derivatives were grown as described in
101 the "Cell lines and growth conditions" section of Materials and methods in [3].
102

103 Bioinformatics analysis

104 Bin-based Analysis of Insertional Mutagenesis Screens (BAIMS)

105 Genetic screens were conducted as described in the "Reporter-based forward genetic
106 screens" section of Materials and methods in [3], except that GT integrations were mapped as
107 follows. FASTQ files containing 36 bp sequencing reads corresponding to genomic sequences
108 flanking retroviral integration sites in both the sorted and unsorted control cells were obtained for
109 the various genetic screens described (National Center for Biotechnology Information (NCBI)
110 Sequence Read Archive (SRA) Study accession number SRP094861). Reads were aligned to the
111 human genome version "GRCh38" using Bowtie alignment software, version 1.0.1 [6], allowing

112 up to 3 base pair mismatches and only reads that aligned to a single locus of the human genome
113 were considered for downstream analysis. The orientation of the reads relative to the "+" or "-"
114 strand of the chromosome, as defined in human genome version GRCh38, was noted.

115 Next, the genome was divided into contiguous, non-overlapping intervals of arbitrary
116 length (250-1000 bp as indicated in the Results and figure legends), which are referred to as
117 "bins", regardless of the location of genes and other genetic elements. Each bin was annotated
118 with any overlapping genes and corresponding features (5'UTR, CDS, intron, and 3'UTR),
119 according to the RefSeq annotations from the University of California, Santa Cruz Table
120 Browser [7] for the GRCh38 assembly of the human genome. An additional genetic feature that
121 we defined as "promoter," encompassing the 2000 bp directly upstream of the TSS of every
122 gene, was also used to annotate any overlapping bins. The orientation of each genetic feature
123 with respect to the chromosome (whether it resides on the "+" or "-" strand of the chromosome,
124 as specified by the RefSeq annotation) was also noted.

125 Each GT insertion considered for downstream analysis was mapped to the bin that
126 encompassed its location in the genome. For each bin, we tallied the number of insertions that
127 mapped to the "+" and to the "-" chromosome strand. This enabled us to determine the number
128 of sense and antisense insertions relative to any genetic feature. For example, a GT insertion that
129 aligned to the "+" chromosome strand was considered to be in the sense orientation with respect
130 to a genetic feature that resided on the "+" chromosome strand, whereas an insertion that aligned
131 to the "-" chromosome strand was considered to be in the antisense orientation with respect to
132 the same genetic feature. Histograms depicting the orientation of insertions across genomic
133 regions or genes of interest could then be generated using insertion counts from the bins
134 contained within the region of interest.

135

136 **Gene-based insertion enrichment analysis**

137 To determine which genes were enriched for total GT insertions in the sorted versus the
138 unsorted cells, all insertions in bins annotated with a given gene and its associated promoter as
139 defined above were aggregated separately for the sorted and unsorted cell populations. Thus, the
140 sum of insertions for a specific gene included both sense and antisense insertions that overlapped
141 with the gene's features, including the promoter. For each gene, a *p*-value for the significance of
142 enrichment was calculated using a one-sided Fisher's exact test run using the "scipy" package
143 (version 0.7.2) in Python 2.7.5 by comparing the frequency of insertions in the gene in the sorted
144 cells to the frequency of insertions in the gene in the unsorted cells; this *p*-value was then
145 corrected for false-discovery rate. Genes were ranked in ascending order based on FDR-
146 corrected *p*-value.

147

148 **Antisense intronic insertion enrichment analysis**

149 This analysis included bins annotated exclusively as intron and containing at least one
150 GT insertion in the antisense orientation with respect to the gene in the sorted cells. An FDR-
151 corrected *p*-value for the significance of antisense insertion enrichment in each of these bins was
152 determined using a one-sided Fisher's exact test (from the "scipy" package for Python)
153 comparing the frequency of antisense insertions in the bin for the sorted versus the unsorted
154 cells. Bins were then ranked in ascending order based FDR-corrected *p*-value (Figs 2B and 2C,
155 S1 File).

156

157 **Upstream insertion enrichment analysis**

158 This analysis included bins annotated exclusively as promoter and containing at least one
159 GT insertion regardless of orientation in the sorted cells. An FDR-corrected *p*-value for the
160 significance of insertion enrichment in each of these bins was determined using a one-sided
161 Fisher's exact test (from "scipy" package for Python) comparing the frequency of insertions in
162 the bin for the sorted versus the unsorted cells. Bins were then ranked in ascending order based
163 on FDR-corrected *p*-value (Figs 2D and 2E, S1 File).
164

165 **Inactivating insertion enrichment analysis**

166 This analysis included bins annotated with any exonic feature (5'UTR, CDS, 3'UTR) and
167 containing at least one GT insertion regardless of orientation in the sorted cells, as well as bins
168 annotated exclusively with intron and containing at least one GT insertion in the sense
169 orientation with respect to the gene in the sorted cells. An FDR-corrected *p*-value for the
170 significance of inactivating insertion (all insertions in bins annotated with 5'UTR, CDS, or
171 3'UTR and only sense insertions in bins annotated exclusively with intron) enrichment in the bin
172 was determined using a one-sided Fisher's exact test (from "scipy" package for Python)
173 comparing the frequency of insertions in the bin for the sorted versus the unsorted cells. Bins
174 were ranked in ascending order based on FDR-corrected *p*-value (Figs 2F and 2G, S1 File).
175

176 **BAIMS pipeline code**

177 The BAIMS pipeline code used for the bioinformatics analysis is available through
178 Github (https://github.com/RohatgiLab/BAIMS-Pipeline).
179

180 **Isolation of cell lines containing GT insertions**

181 All clonal cell lines containing specified GT insertions were isolated as described in the
182 "Isolation of APC^{KO-2} mutant cell line containing a GT insertion" section of Materials and
183 methods in [3].
184 The $TFAP4^{GT}$ cell line containing an antisense GT insertion in the first intron of *TFAP4*
185 was isolated from the WNT positive regulator high stringency screen using the reverse primer
186 Wntlow TFAP4 AS II (5'-GCTGCACACGTGTAGACACTC-3').
187 $LRP6^{GT}$-1(Up) and $LRP6^{GT}$-2(Up) cell lines, containing antisense GT insertions upstream
188 of the *LRP6* TSS, and the $LRP6^{GT}$-3(Int) cell line, containing a sense GT insertion in the first
189 intron of *LRP6*, were isolated from the WNT positive regulator high stringency screen using the
190 reverse primers LRP6UP-ASGT-Loc-2 (5'-GCAGTGTGTAATATCTCATTCCC-3'), LRP6UP-
191 ASGT-Loc-1 (5'-GGAGACTCCCATTACTCTCTGTT-3') and Wntlow LRP6 (5'-
192 TGTGGGAAAACTTTGTAATATGC-3'), respectively.
193 The genomic location of the GT insertion in each isolated cell line is indicated in S5 File.
194

195 ## Analysis of WNT reporter fluorescence

196 WNT reporter fluorescence (Figs 3B and 4C) was measured in WT HAP1-7TGP cells or
197 derivatives thereof as described in the "Analysis of WNT reporter fluorescence" section of
198 Materials and methods of [3].
199

200 ## Quantitative RT-PCR (qPCR) analysis

201 All mRNA measurements were made as described in the "Quantitative RT-PCR analysis"
202 section of Materials and methods in [3] using the *AXIN2* and *HPRT1* primers described therein
203 (Figs 3C, 3D, 4D, and 4G), the following forward and reverse primers for *TFAP4* (Fig 3D):

204 hTFAP4-RT-PCR-1-FOR (5'-GAGGGCTCTGTAGCCTTGC-3') and hTFAP4-RT-PCR-1-REV
205 (5'-GAATCCCGCGTTGATGCTCT-3'), and the following forward and reverse primers
206 spanning two pairs of contiguous exons for *LRP6* (Fig 4G): qPCR-LRP6-Exons-1-2-For (5'-
207 GCTTCTGTGTGCTCCTGAG-3'), qPCR-LRP6-Exons-1-2-Rev (5'-
208 TCCAAGCCTCCAACTACAATC-3'), qPCR-LRP6-Exons-7-8-For (5'-
209 GGAGATGCCAAAACAGACAAG -3'), and qPCR-LRP6-Exons-7-8-Rev (5'-
210 CAGTCCAGTAAACATAGTCACCC -3').
211

212 # Immunoblot analysis of WT HAP1-7TGP and mutant cell lines

213 ## Immunoblot analysis of TFAP4 (Fig 3E)

214 This analysis was performed as described in the "Immunoblot analysis of HAP1-7TGP
215 and mutant cell lines - Immunoblot analysis of total AXIN1 and AXIN2" section of Materials
216 and methods in [3] with some modifications. Cell pellets harvested from confluent 6 cm dishes
217 were resuspended in 100 µl of ice-cold RIPA lysis buffer (50 mM Tris-HCl pH 8.0, 150 mM
218 NaCl, 2% NP-40, 0.25% deoxycholate, 0.1% SDS, 1X SIGMA*FAST* protease inhibitors, 10%
219 glycerol), sonicated in a Bioruptor 300 (Diagenode) 2 x 30 sec in the high setting, centrifuged 10
220 min at 20,000 x g, and the supernatant was recovered.
221 The protein concentration in the supernatant was quantified using the Pierce BCA Protein
222 Assay Kit. Samples were normalized by dilution with RIPA lysis buffer, further diluted with 4X
223 LDS sample buffer supplemented with 50 mM TCEP, heated for 5 min at 95°C, and 40 µg of
224 total protein were electrophoresed alongside Precision Plus Protein All Blue Prestained Protein
225 Standards in NuPAGE 4-12% Bis-Tris gels using 1X NuPAGE MOPS SDS running buffer.

226 Proteins were transferred to nitrocellulose membranes using 1X NuPAGE transfer buffer
227 + 10% methanol, membranes were stained with Ponceau S to assess loading, washed and
228 blocked with TBST + 5% skim milk. Blots were incubated with rabbit anti-AP4 (TFAP4),
229 washed with TBST, incubated with Peroxidase AffiniPure anti-rabbit secondary antibody,
230 washed with TBST followed by TBS, and developed with SuperSignal™ West Pico
231 Chemiluminescent Substrate and SuperSignal West Femto Maximum Sensitivity Substrate
232 (Thermo Fisher Scientific Cat. # 34080 and 34095).
233

234 **Immunoblot analysis of total LRP6 (Fig 4E and C in S4 Fig)**

235 This analysis was performed as described in the in the previous section. 75 µg of total
236 protein were loaded in duplicate and electrophoresed in a NuPAGE 4-12% Bis-Tris gel.
237 Following the transfer step, the nitrocellulose membrane was cut between the 50 and 75 kDa
238 molecular weight standards and blocked for 1 hour with Odyssey Blocking Buffer. The top
239 membrane was incubated with rabbit anti-LRP6 primary antibody, and the bottom membrane
240 was incubated with mouse anti-ACTIN primary antibody. Membranes were washed with TBST
241 and incubated with IRDye 800CW donkey anti-rabbit IgG and IRDye 800CW donkey anti-
242 mouse IgG secondary antibodies, respectively. Membranes were washed with TBST followed by
243 TBS, and imaged using the Li-Cor Odyssey imaging system. Acquisition parameters in the
244 manufacturer's Li-Cor Odyssey Image Studio software were set so as to avoid saturated pixels in
245 the bands of interest, and bands were quantified using manual background subtraction. The
246 integrated intensity for LRP6 was normalized to that for ACTIN in the same lane and the average
247 +/- SD from duplicate lanes was used to represent the data in Fig 4E.
248

249 **LRP6 cell-surface staining of WT HAP1-7TGP and mutant cell lines**

250 **(Fig 4F)**

251 Approximately 24 hr before staining, cells were seeded in a 6-well plate at a density of
252 $2x10^5$ per well and grown in 2.5 ml of CGM 2 (defined in [3]). On the following day the cells
253 were washed once in 3 ml of phosphate buffered saline (PBS), harvested in 0.5 ml of Accutase
254 Cell Detachment Reagent (Thermo Fisher Scientific Cat. # NC9839010), resuspended in 1.5 ml
255 of ice-cold CGM 2 and centrifuged 4 min at 400 x g at 4°C (all subsequent centrifugation steps
256 were done in the same way). Cells were washed with 2 ml of ice-cold Iscove's Modified
257 Dulbecco's Medium (IMDM) with L-glutamine, with HEPES, without Alpha-Thioglycerol (GE
258 Healthcare Life Sciences Cat. # SH30228.01) + 1% Fetal Bovine Serum (FBS) (Atlanta
259 Biologicals Cat. # S11150), centrifuged and resuspended in 150 µl of mouse anti-LRP6 primary
260 antibody in IMDM + 1% FBS. Following a 1 hr incubation on ice, cells were washed with 1.8 ml
261 of ice-cold IMDM + 1% FBS, centrifuged, washed with 2 ml of ice-cold IMDM +1% FBS and
262 centrifuged again. Cells were resuspended in 150 µl of secondary antibody in ice-cold IMDM +
263 1% FBS and incubated on ice for 30 minutes. Cells were washed with 1.8 ml of ice-cold IMDM
264 + 1% FBS, centrifuged, washed with 2 ml of ice-cold IMDM + 1% FBS and centrifuged again.
265 Cells were resuspended in 200 µl of PBS + 2% FBS and LRP6 cell-surface fluorescence was
266 analyzed on a BD Accuri RUO Special Order System (BD Biosciences).

Results

Bin-based Analysis of Insertional Mutagenesis Screens (BAIMS)

Haploid genetic screens rely on the phenotypic selection of a population of cells mutagenized by integration of a GT-bearing retrovirus. GTs, which contain a splice acceptor (SA) and a transcriptional termination polyadenylation signal (pA), can disrupt protein-coding genes in two ways: (1) by inserting into an exon in either the sense or antisense orientation relative to the coding sequence of the gene or (2) by inserting into an intron in the sense orientation, such that the directional SA causes the GT to be spliced to the 3'-end of the preceding exon, resulting in a transcript that undergoes premature termination (A-D in S2 Fig). Indeed, most hits in haploid genetic screens exhibit a bias towards such inactivating sense insertions in introns [8]. GT insertions can theoretically also perturb gene regulation by directly interrupting a regulatory protein-binding site on DNA, by terminating a regulatory transcript, or by altering chromatin structure. Therefore, in principle it should be possible to find GT insertion patterns indicative of such regulatory mechanisms.

In order to map GT insertions in a way that would enable us to identify regulatory elements, we devised a bioinformatics pipeline that was completely agnostic to the boundaries of annotated genes and simply tracked the number and orientation of GT insertions in short genomic intervals of arbitrary size, defined as "bins" (Fig 1A). We refer to this approach as "Bin-based Analysis of Insertional Mutagenesis Screens", or BAIMS. Sequencing reads adjacent to the location of GT insertions found in sorted (phenotypically selected) and unsorted (control) cells from haploid genetic screens were aligned to the human genome and assigned to the bin that encompassed the insertion (Fig 1B). The orientation of each insertion on the chromosome

290 was defined according to the GRCh38 assembly of the human genome. Each bin was also
291 annotated with any relevant genetic features it overlapped with (5' untranslated region (5'UTR),
292 coding domain sequence (CDS), intron and 3' untranslated region (3'UTR)), using the RefSeq
293 annotations from the University of California, Santa Cruz Table Browser [7] for the GRCh38
294 assembly of the human genome. We also defined an additional genetic feature, designated
295 "promoter", as the 2000 base pairs (bp) upstream of the TSS of each gene. This region typically
296 includes the minimal promoter but may also contain other cis-regulatory elements. We annotated
297 bins overlapping with this feature accordingly. The relative orientation of any insertion with
298 respect to any feature can therefore be determined, allowing us to observe patterns of sense and
299 antisense GT insertions across features of interest (Fig 1C). This information can be displayed in
300 a histogram depicting insertions over any genomic region of interest (Fig 1C), providing a high-
301 resolution picture of the insertional landscape. Thus, BAIMS enables us to identify atypical
302 patterns of GT insertions in specific genetic features that could be indicative of regulatory
303 elements.
304
305 **Fig 1. Schematic of Bin-based Analysis of Insertional Mutagenesis Screens (BAIMS).**
306 (A) The human genome is computationally divided into "bins" (pictured as rectangles with black
307 dotted lines), which encompass contiguous segments of DNA of an equal arbitrary length.
308 Throughout this study, we used bins of 250 bp or 1000 bp in length, depending on the resolution
309 required for the analysis. The boundaries of annotated genetic features, including genes and
310 regulatory elements, are ignored. The depicted fictitious gene is modeled after a RefSeq gene
311 track following the University of California, Santa Cruz (UCSC) genome browser display
312 conventions: coding exons are represented by tall blocks, UTRs by shorter blocks, and introns by
313 horizontal lines connecting the blocks. The arrow indicates the gene's TSS. (B) Sequencing

314 reads flanking the location of individual GT insertions in the control and selected cell
315 populations from a haploid genetic screen are mapped to the human genome and assigned to the
316 bin that encompasses the location of the insertion. The orientation of each insertion relative to
317 the chromosome is noted. Bins are also annotated with any overlapping genetic features. These
318 include promoter (defined as the 2000 bp upstream of the TSS, indicated by a horizontal dotted
319 line), 5'UTR, CDS, intron, and 3'UTR. The orientation of the feature relative to the chromosome
320 is also noted. (C) For the bin-based analysis, the number and orientation of GT insertions in
321 consecutive bins along any defined portion of the genome (including but not limited to genes) is
322 determined and can be depicted in a histogram (the number of sense GT insertions per bin is
323 arbitrarily shown above the horizontal line labeled "0", and the number of antisense insertions
324 below), enabling the visualization of insertion patterns at sub-gene resolution. (D) For the gene-
325 based analysis, GT insertions in bins that overlap with genes can be summed to obtain a total
326 insertion count for each gene. The significance of GT enrichment for every gene is calculated by
327 comparing the total number of insertions per gene found in the selected versus the control cell
328 populations (see Materials and Methods for details).
329
330 The overall enrichment of GT insertions for any given gene in the selected versus the
331 control cells from a haploid genetic screen can also be assessed by aggregating the insertions
332 found in all bins that overlap with the gene (Fig 1D; see Materials and Methods). We refer to this
333 analysis, which produces a significance score for GT enrichment comparable to that of previous
334 analyses [3], as "gene-based insertion enrichment analysis".
335

Identification of regulatory elements through the analysis of bins with atypical GT insertion patterns

Our previous analysis [3] focused on GT insertions predicted to inactivate protein-coding genes and non-coding RNAs as outlined above: sense and antisense insertions in exons, and sense insertions in introns (B-D in S2 Fig). To identify regulatory elements, we searched for GT insertion patterns distinct from these. Because the GT retrovirus has a strong propensity to integrate near TSSs, promoters and enhancers, we limited our analysis to non-coding regions within and adjacent to genes. We used BAIMS to look for enrichment of antisense insertions in introns, which would not be predicted to interrupt protein-coding transcripts (E in S2 Fig), and for enrichment of insertions in either orientation in the regions upstream of the TSS of genes (F and G in S2 Fig). Since each bin is annotated with the genetic features it overlaps with (Fig 2A), we could readily identify these distinct patterns of GT insertions.

Fig 2. BAIMS identifies atypical GT insertion patterns in screens for regulators of WNT signaling.

(A) Schematic depicting various patterns of GT insertions relative to genetic features in the containing bins, used for the antisense intronic, upstream and inactivating insertion enrichment analyses (see text for details). A fictitious gene modeled after a RefSeq gene track, with GT insertions in the sense orientation relative to the gene depicted above the track and in the antisense orientation depicted below it. The antisense intronic insertion enrichment analysis accounts for antisense GT insertions in bins annotated exclusively as intron (depicted in blue) and the upstream insertion enrichment analysis accounts for both sense and antisense insertions in bins annotated exclusively as promoter (depicted in red). These two classes of insertions had

359 been ignored in previous gene-based analyses of haploid genetic screens [3]. The inactivating
360 insertion enrichment analysis accounts for both sense and antisense insertions in bins annotated
361 as 5'UTR, CDS, or 3'UTR, as well as sense insertions in bins annotated exclusively as intron;
362 these insertions (depicted in black) include all the gene-inactivating insertions used in previous
363 analyses. (B-G) Circle plots depicting the results of antisense intronic (B, C), upstream (D, E),
364 and inactivating (F, G) insertion enrichment analyses for the WNT positive regulator high
365 stringency (B, D, and F) and low stringency (C, E, and G) screens. Circles represent individual
366 1000 bp bins. The y-axis indicates the significance of GT insertion enrichment in the sorted
367 versus the unsorted, control cells, expressed in units of $-\log_{10}$(FDR-corrected *p*-value), and the x-
368 axis indicates the 5000 bins with the smallest FDR-corrected *p*-values, arranged in random order.
369 Circles representing bins with an FDR-corrected *p*-value < 0.01 are colored and labeled with the
370 name of the gene with which the bin overlaps. Circles representing bins corresponding to the
371 same gene are depicted in the same color. The diameter of each circle is proportional to the
372 number of independent GT insertions mapped to the corresponding bin in the sorted cells, which
373 is also indicated next to the gene name for enriched bins.
374
375 To identify regulatory elements in introns, we looked for enrichment of antisense
376 insertions in bins exclusively annotated as intron (Fig 2A); we refer to this analysis as "antisense
377 intronic insertion enrichment analysis." To identify regulatory elements in regions immediately
378 upstream of genes, we looked for enrichment of both sense and antisense GT insertions in bins
379 exclusively annotated as promoter (Fig 2A); we refer to this analysis as "upstream insertion
380 enrichment analysis." To distinguish features identified in the previous two analyses from the
381 more typical disruption of protein-coding genes or non-coding RNAs by GT insertions, we

382 looked for enrichment of gene-inactivating insertions, as defined above (sense and antisense
383 insertions in bins annotated with 5'UTR, CDS or 3'UTR, and sense insertions in bins annotated
384 exclusively as intron; see Fig 2A); we refer to this analysis as "inactivating insertion enrichment
385 analysis."

386 These three analyses were applied to the data from two screens for positive regulators of
387 signaling following stimulation with WNT3A, henceforth referred to as the WNT positive
388 regulator high and low stringency screens, which differed only in the stringency of selection [3].
389 In these screens, HAP1 cells harboring a WNT-responsive GFP reporter (hereafter referred to as
390 "WT HAP1-7TGP") were mutagenized with GT retrovirus, treated with WNT3A and sorted by
391 fluorescence activated cell sorting (FACS) for cells that exhibited the lowest 2% (high
392 stringency) or the lowest 10% (low stringency) signaling activity. These screens enabled us to
393 identify known and new regulators in the WNT pathway [3].

394 Antisense intronic insertion enrichment analysis of the WNT positive regulator high and
395 low stringency screens produced only one significant (FDR-corrected p-value < 0.01) bin (Figs
396 2B and 2C, S1 File), which mapped to the gene *TFAP4*, one of the top hits from these screens
397 [3]. Upstream insertion enrichment analysis of the same screens produced only one significant
398 bin upstream of *LRP6* (Figs 2D and 2E, S1 File), which was the top hit of both of these screens
399 [3]. These results are markedly different from those of the inactivating insertion enrichment
400 analysis of the same screens (Figs 2F and 2G, S1 File), which revealed bins in many of the same
401 genes identified as significant hits in these screens [3].

402 In the sections that follow, we tested if the GT insertion patterns identified in *TFAP4* and
403 *LRP6* by the antisense intronic insertion and upstream insertion enrichment analyses,
404 respectively, reflected regulatory effects on gene expression.

405

406 Antisense GT insertions in the first intron of *TFAP4* disrupt gene 407 expression

408 The second top hit in the WNT positive regulator high and low stringency screens was 409 *TFAP4*, encoding the transcription factor TFAP4, which we have shown to be a positive 410 regulator of the WNT pathway acting downstream of the key transcriptional co-activator β- 411 catenin (CTNNB1) [3]. As is common for top hits of haploid genetic screens, the 5' end of 412 *TFAP4* was significantly enriched for inactivating GT insertions, including many sense and 413 antisense insertions in the first exon as well as sense insertions in the first intron, which are all 414 expected to disrupt the *TFAP4* coding sequence (Fig 3A and A in S3 Fig). However, the single 415 bin identified in the antisense intronic insertion enrichment analysis (Figs 2B and 2C, S1 File) 416 was also located in the first intron and it contained a comparable number of antisense GT 417 insertions (Fig 3A and A in S3 Fig), which would not be predicted to disrupt the *TFAP4* coding 418 sequence. This pattern of GT insertion enrichment was not seen for *TFAP4* in the mutagenized 419 but unsorted cells used as a control for the WNT positive regulator screens (Fig 3A) or for other 420 top hits, such as *DOT1L*, in the sorted cells from these same screens (B in S3 Fig). These results 421 suggested that antisense GT insertions in the first intron of *TFAP4* (which would not be 422 predicted to terminate the *TFAP4* transcript) reduced WNT signaling.

423

424 **Fig 3. Antisense GT insertions in the first intron of *TFAP4* disrupt gene expression and 425 impair WNT signaling.**

426 (A) The histogram indicates the number and orientation of GT insertions mapped to *TFAP4* in 427 unsorted cells and in the sorted cells from the WNT positive regulator low stringency screen.

428 Values above the horizontal line labeled "0" indicate sense insertions relative to the coding
429 sequence of the gene, and values below it indicate antisense insertions. The x-axis represents
430 contiguous 250 bp bins to which insertions were mapped (Chromosome 16, 4257249-4273000
431 bp). Insertions mapped for the different cell populations indicated in the legend are depicted by
432 traces of different colors. A RefSeq gene track for *TFAP4* (following UCSC genome browser
433 display conventions, described in the legend of Fig 1A) and an ENCODE track for histone3-
434 lysine27-acetylation, a marker for enhancer activity (taken from the UCSC genome browser), are
435 shown underneath the graph. The black rectangle above the gene track indicates the location of
436 the bin identified in the antisense intronic insertion enrichment analyses of both the WNT
437 positive regulator low stringency and high stringency screens. The black star denotes the position
438 of the antisense GT insertion in the $TFAP4^{GT}$ clonal cell line used for further characterization. A
439 scale bar is provided beneath the gene track for reference. (B) Fold-induction in WNT reporter
440 (median +/- standard error of the median (SEM) EGFP fluorescence from 10,000 cells) following
441 treatment with 50% WNT3A conditioned media (CM). (C) *AXIN2* mRNA (average +/- standard
442 deviation (SD) of *AXIN2* mRNA normalized to *HPRT1* mRNA, each measured in triplicate
443 qPCR reactions) relative to untreated cells. Where indicated, cells were treated with 50%
444 WNT3A CM. (D) *TFAP4* mRNA (average +/- SD of *TFAP4* mRNA normalized to *HPRT1*
445 mRNA, each measured in triplicate qPCR reactions) relative to WT HAP1-7TGP cells. (E)
446 Immunoblot of TFAP4. The middle panel shows a higher exposure of the same blot shown in the
447 top panel, and the bottom panel displays Ponceau S staining of the same blot as a loading
448 control. Molecular weight standards in kiloDaltons (kDa) are indicated to the left of each blot.
449
450 To confirm this prediction, we isolated a clonal cell line harboring an antisense GT

451 insertion in the first intron of *TFAP4* (we designate this cell line TFAP4^{GT}; see Fig 3A and
452 Materials and Methods). WNT3A-induced reporter activation was nearly eliminated in TFAP4^{GT}
453 cells when compared to WT HAP1-7TGP cells (Fig 3B). Expression of *AXIN2* mRNA, a
454 universal target gene of the pathway, following treatment with WNT3A was also reduced
455 substantially in TFAP4^{GT} cells (Fig 3C). Given its location within the boundaries of the *TFAP4*
456 gene, we tested whether the antisense GT insertion affected expression of *TFAP4* itself. Both
457 *TFAP4* mRNA and protein levels were severely reduced in TFAP4^{GT} cells, explaining the
458 observed defect in pathway activity (Figs 3D and 3E). A higher exposure of the TFAP4
459 immunoblot from TFAP4^{GT} cells revealed a faint band corresponding to TFAP4 (Fig 3E),
460 indicating that the antisense GT insertion in the first intron of *TFAP4* reduced expression of a
461 full-length transcript and protein as opposed to disrupting the coding sequence.
462

463 Antisense GT insertions upstream of *LRP6* reduce LRP6 protein
464 abundance independently of mRNA levels

465 *LRP6* encodes a required co-receptor for WNT ligands and was the top hit of the WNT
466 positive regulator high and low stringency screens [3]. As expected, most GT insertions in the
467 *LRP6* gene proper (downstream of the TSS) were in the sense orientation with respect to the
468 coding sequence (Figs 4A and 4B, and A and B in S4 Fig). However, our upstream insertion
469 enrichment analysis also revealed a bin enriched for GT insertions located upstream of the TSS
470 (Figs 2D and 2E, S1 File). A closer inspection of the region surrounding this bin revealed a
471 pronounced enrichment of antisense insertions extending from about 1 to 3.5 kilobase pairs (kbp)
472 upstream of the TSS (Fig 4B and B in S4 Fig). Importantly, this region was located upstream of
473 the annotated *LRP6* promoter in *Ensembl* (Fig 4B). These GT insertion patterns were not

474 observed in the mutagenized but unsorted cells used as a control for the WNT positive regulator
475 screens (Figs 4A and 4B). These results suggested that antisense insertions upstream of *LRP6*
476 impaired WNT signaling.
477
478 **Fig 4. Antisense GT insertions upstream of *LRP6* reduce LRP6 protein expression and**
479 **impair WNT signaling.**
480 (A) The histogram indicates the number and orientation of GT insertions mapped to *LRP6* and to
481 the region ~12.5 kbp upstream of the TSS in unsorted cells and in the sorted cells from the WNT
482 positive regulator low stringency screen. See legend to Fig 3A for details. The x-axis represents
483 contiguous 250 bp bins to which insertions were mapped (Chromosome 12, 12116000-12279249
484 bp). (B) The histogram shows an expanded view of the 5' end of *LRP6* and the region ~12.5 kbp
485 upstream of the TSS (left of the vertical dotted line), with traces for GT insertions mapped in
486 unsorted cells and in the sorted cells from the WNT positive regulator low stringency screen. The
487 x-axis represents contiguous 250 bp bins to which insertions were mapped (Chromosome 12,
488 12262500-12279249 bp). The green rectangle above the gene track indicates the location of the
489 *LRP6* promoter according to *Ensembl* and the black rectangle indicates the location of the bin
490 identified in the upstream insertion enrichment analyses of the WNT positive regulator low
491 stringency and high stringency screens. The black and red stars denote the position of the
492 antisense GT insertions in the LRP6^{GT}-1(Up) and LRP6^{GT}-2(Up) clonal cell lines, respectively,
493 and the blue star denotes the position of the sense GT insertion in the LRP6^{GT}-3(Int) cell line. (C)
494 Fold-induction in WNT reporter (median +/- SEM EGFP fluorescence from 20,000 cells)
495 following treatment with 50% WNT3A CM. (D) Fold-induction in *AXIN2* mRNA (average +/-
496 SD of *AXIN2* mRNA normalized to *HPRT1* mRNA, each measured in triplicate qPCR reactions)
497 following treatment with 50% WNT3A CM. (E) Quantification of immunoblot analysis of total

498 LRP6 protein (average +/- SD LRP6 intensity normalized to ACTIN intensity from samples run
499 in duplicate) shown as percentage of WT HAP1-7TGP. The blot used for quantification is shown
500 in C in S4 Fig. (F) Cell surface LRP6 protein (median +/- SEM cell surface LRP6
501 immunofluorescence from 20,000 cells) shown as percentage of WT HAP1-7TGP. (G) *LRP6*
502 mRNA (average +/- SD of *LRP6* mRNA, measured using two different primer pairs, normalized
503 to *HPRT1* mRNA, each measured in triplicate qPCR reactions) shown relative to WT HAP1-
504 7TGP cells.
505
506 To test this possibility, we isolated two clonal cell lines containing antisense GT
507 insertions in the region upstream of *LRP6* (we designate these cell lines LRP6^{GT}-1(Up) and
508 LRP6^{GT}-2(Up); see Fig 4B and Materials and Methods) and as a control we isolated a clonal cell
509 line with a sense GT insertion in the first intron of *LRP6* that is predicted to disrupt the *LRP6*
510 coding sequence (we designate this cell line LRP6^{GT}-3(Int); see Fig 4B and Materials and
511 Methods). Both LRP6^{GT}-1(Up) and LRP6^{GT}-2(Up) cells demonstrated significantly reduced
512 WNT reporter activation and *AXIN2* mRNA accumulation following treatment with WNT3A
513 when compared to WT HAP1-7TGP cells (Figs 4C and 4D). The most plausible explanation for
514 how the GT insertions reduced WNT signaling would be down-regulation of LRP6, which is
515 indeed what we observed when we measured total and cell-surface levels of LRP6 protein.
516 LRP6^{GT}-1(Up) and LRP6^{GT}-2(Up) cells exhibited a 75-84% reduction in total LRP6 protein and
517 a 68-71% reduction in cell-surface LRP6 compared to WT cells (Figs 4E and 4F). LRP6^{GT}-3(Int)
518 cells exhibited greater, >99% and 94% reductions in total and cell-surface LRP6, respectively,
519 compared to WT cells, as would be expected from the disruption of the *LRP6* coding sequence
520 caused by the sense GT insertion in the first intron (Figs 4E and 4F).

521 Unexpectedly, despite the reduction in LRP6 protein observed in LRP6^{GT}-1(Up) and
522 LRP6^{GT}-2(Up) cells harboring antisense GT insertions upstream of the *LRP6* promoter, we did
523 not observe a corresponding decrease in *LRP6* mRNA (Fig 4G). In an important control, *LRP6*
524 mRNA levels were indeed markedly reduced in LRP6^{GT}-3(Int) cells carrying a sense intronic GT
525 insertion that disrupts the coding sequence (Fig 4G). These results suggest that antisense GT
526 insertions upstream of *LRP6* diminished signaling by an enigmatic mechanism that reduced
527 LRP6 protein levels without altering mRNA levels, rather than by simply disrupting the *LRP6*
528 promoter. Interestingly, sequence elements with similar properties have been described upstream
529 of promoter elements for heat shock target genes in yeast [9].

530 # Discussion

531 We developed a new bioinformatics tool to analyze haploid genetic screens with the
532 explicit goal of uncovering regulatory elements. We analyzed screen data in a way that discerned
533 GT insertion patterns distinct from those predicted to disrupt the coding sequence of genes, and
534 found that atypical insertions in introns and regions upstream of the TSS can cause down-
535 regulation of genes, leading to the phenotype selected for during the screen. When we applied
536 this new analysis to haploid genetic screens interrogating the WNT pathway, we found that
537 antisense GT insertions in the first intron of *TFAP4* and upstream of the *LRP6* promoter resulted
538 in marked changes in the expression of these genes. These types of insertions had not been
539 accounted for in previous analyses of haploid genetic screens.

540 The identified GT insertions could disrupt regulatory elements such as promoters,
541 enhancers, antisense transcripts or splicing sequences. In the case of *TFAP4*, most of the
542 insertions were located in the first intron and overlapped with a strong enhancer signal (Fig 3A),
543 suggesting they may disrupt an enhancer. Previous studies have shown that *TFAP4* is directly
544 regulated by c-MYC and that the first intron of *TFAP4* in fact contains four c-MYC binding sites
545 [10, 11], three of which are encompassed by the bin identified in our antisense intronic insertion
546 enrichment analysis (Figs 2B and 2C). In future studies, it will be important to test whether the
547 antisense insertions found in the first intron of *TFAP4* down-regulate *TFAP4* mRNA (Fig 3D) by
548 disrupting c-MYC binding or through an alternative mechanism.

549 Similarly, LRP6 protein was down-regulated in the $LRP6^{GT}$-1(Up) and $LRP6^{GT}$-2(Up)
550 cell lines containing antisense GT insertions upstream of the *LRP6* promoter (Figs 4E and 4F).
551 Surprisingly, *LRP6* mRNA levels were not reduced in these same cell lines, suggesting a
552 mechanism regulating LRP6 protein levels. In yeast, genomic sequences upstream of genes that

553 have no effect on mRNA levels can instead regulate protein levels [9]. The selective enrichment
554 of antisense versus sense GT insertions in the region upstream of the *LRP6* promoter in cells
555 sorted for low WNT reporter fluorescence (Figs 4A and 4B) suggests that such insertions are not
556 merely disrupting an enhancer or promoter. Instead, we speculate that these GT insertions may
557 disrupt an antisense transcript or another directional element residing on the antisense strand that
558 positively regulates *LRP6* expression. Since no such elements have been described, it will be
559 important to elucidate the nature of this regulatory mechanism in future studies.

560 Unlike other more focused efforts to identify regulatory regions associated with a given
561 gene or set of genes [12-18], our untargeted approach enables the genome-wide identification of
562 cis-regulatory elements involved in any phenotype that can be probed through a haploid genetic
563 screen. Identification of such elements may not be feasible with RNA interference-based screens
564 because they require that the target genomic sequences be transcribed. CRISPR-based
565 technologies to screen for regulatory modules on a genome scale are still limited by the focused
566 mutagenesis or transcriptional modulation of predetermined sequences in the genome [19-22].
567 However, focused CRISPR-based approaches would be powerful tools to precisely delineate any
568 regulatory element found though our analysis.

569 While we found new regulatory elements in two central regulators of WNT signaling, we
570 note that our current study is most likely under-powered to comprehensively detect all regulatory
571 elements in the genome affecting the WNT pathway for several reasons. First, we used deep
572 sequencing datasets from previous screens [3] that were designed to uncover protein coding
573 genes involved in WNT signaling. The sequencing depth used to map insertions in these
574 previous screens was sufficient to saturate the protein-coding genome, but is insufficient to
575 interrogate the much larger non-coding genome. Second, the propensity of the retroviral-based

576 mutagen used in this study to insert near TSSs, promoters and enhancers limited our search for
577 regulatory elements to regions within and adjacent to genes. Our methodology could in principle
578 be extended to identify regulatory elements located anywhere in the genome by using agents that
579 integrate in a truly unbiased manner and then exhaustively mapping insertions in both the
580 selected and unselected cell populations by sequencing at greater depth. Finally, because we
581 assigned bins disregarding gene boundaries, our analysis may have missed regulatory elements
582 in bins that overlapped with both an exon and an intron (such bins would have been excluded
583 from the antisense intronic insertion enrichment analysis), and elements in bins that overlapped
584 with features located both upstream and downstream of the TSS (such bins would have been
585 excluded from the upstream insertion enrichment analysis). Reducing the size of the bins could
586 ameliorate this problem, but at the expense of statistical power to determine the significance of
587 GT insertion enrichment due to a reduction in GT insertions per bin and an increase in the
588 multiple testing correction for a larger number of bins. Alternatively, computing GT insertions in
589 intervals defined by the boundaries of genetic features such as introns or promoters (rather than
590 bins of a predetermined size) could also help this issue, but would limit the analysis to annotated
591 regions of the genome.
592 The analysis described in this work provides an untargeted and genome-scale method to
593 identify both genes and regulatory elements involved in any biological process that can be
594 probed by a haploid genetic screen. We hope that this bioinformatics analysis, available through
595 Github (https://github.com/RohatgiLab/BAIMS-Pipeline), empowers other researchers to extract
596 new insights about gene regulation from the growing body of insertional mutagenesis screen
597 data.

Acknowledgements

We thank members of the Rohatgi, Salzman and Carette labs for input on the project.

600 # References

601 1. Consortium EP. An integrated encyclopedia of DNA elements in the human genome.
602 Nature. 2012;489(7414):57-74.
603 2. Wright JB, Sanjana NE. CRISPR Screens to Discover Functional Noncoding Elements.
604 Trends Genet. 2016;32(9):526-9.
605 3. Lebensohn AM, Dubey R, Neitzel LR, Tacchelly-Benites O, Yang E, Marceau CD, et al.
606 Comparative genetic screens in human cells reveal new regulatory mechanisms in WNT
607 signaling. Elife. 2016;5.
608 4. Vrljicak P, Tao S, Varshney GK, Quach HN, Joshi A, LaFave MC, et al. Genome-Wide
609 Analysis of Transposon and Retroviral Insertions Reveals Preferential Integrations in Regions of
610 DNA Flexibility. G3 (Bethesda). 2016;6(4):805-17.
611 5. Egawa T, Littman DR. Transcription factor AP4 modulates reversible and epigenetic
612 silencing of the Cd4 gene. Proc Natl Acad Sci U S A. 2011;108(36):14873-8.
613 6. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment
614 of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
615 7. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The
616 UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32(Database issue):D493-6.
617 8. Carette JE, Guimaraes CP, Wuethrich I, Blomen VA, Varadarajan M, Sun C, et al.
618 Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. Nat
619 Biotechnol. 2011;29(6):542-6.
620 9. Zid BM, O'Shea EK. Promoter sequences direct cytoplasmic localization and translation
621 of mRNAs during starvation in yeast. Nature. 2014;514(7520):117-21.
622 10. Jung P, Hermeking H. The c-MYC-AP4-p21 cascade. Cell Cycle. 2009;8(7):982-9.

623 11. Jung P, Menssen A, Mayr D, Hermeking H. AP4 encodes a c-MYC-inducible repressor
624 of p21. Proc Natl Acad Sci U S A. 2008;105(39):15046-51.
625 12. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al.
626 Systematic mapping of functional enhancer-promoter connections with CRISPR interference.
627 Science. 2016;354(6313):769-73.
628 13. Korkmaz G, Lopes R, Ugalde AP, Nevedomskaya E, Han R, Myacheva K, et al.
629 Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat
630 Biotechnol. 2016;34(2):192-8.
631 14. Smeenk L, van Heeringen SJ, Koeppel M, van Driel MA, Bartels SJ, Akkers RC, et al.
632 Characterization of genome-wide p53-binding sites upon stress response. Nucleic Acids Res.
633 2008;36(11):3639-54.
634 15. Canver MC, Lessard S, Pinello L, Wu Y, Ilboudo Y, Stern EN, et al. Variant-aware
635 saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-
636 associated loci. Nat Genet. 2017.
637 16. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, et al. BCL11A
638 enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature.
639 2015;527(7577):192-7.
640 17. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, et al. High-
641 throughput mapping of regulatory DNA. Nat Biotechnol. 2016;34(2):167-74.
642 18. Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, Joung J, et al. High-resolution
643 interrogation of functional elements in the noncoding genome. Science. 2016;353(6307):1545-9.
644 19. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference
645 (CRISPRi) for sequence-specific control of gene expression. Nat Protoc. 2013;8(11):2180-96.

646 20. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing
647 CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell.
648 2013;152(5):1173-83.
649 21. Perez-Pinera P, Kocak DD, Vockley CM, Adler AF, Kabadi AM, Polstein LR, et al.
650 RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat Methods.
651 2013;10(10):973-6.
652 22. Kearns NA, Genga RM, Enuameh MS, Garber M, Wolfe SA, Maehr R. Cas9 effector-
653 mediated regulation of transcription and differentiation in human pluripotent stem cells.
654 Development. 2014;141(1):219-23.

655 **S1 File. BAIMS output.**

656 Ranked lists of bins from the bin-based analyses. Each sheet of the Excel file contains a ranked
657 list of bins determined by either the antisense intronic, upstream, or inactivating insertion
658 enrichment analysis applied to either the WNT positive regulator low stringency or high
659 stringency screen. The screen and type of bin-based analysis is indicated at the top of every
660 sheet. The location of each bin in the human genome, the genes overlapping with the bin, and the
661 FDR-corrected *p*-values generated by the bin-based analysis are specified. For each bin, the
662 number of antisense intronic insertions, upstream insertions (sum of sense and antisense
663 insertions), or inactivating insertions (sum of sense and antisense insertions for bins overlapping
664 with a 5'UTR, CDS, or 3'UTR, or sense insertions only for bins overlapping exclusively with an
665 intron) found within the bin in unsorted (control) and sorted cells are also indicated. The total
666 number of insertions mapped in the unsorted cells and in the sorted cells are also shown.

667 **S2 Fig. Possible outcomes of GT insertions in different genetic features.**

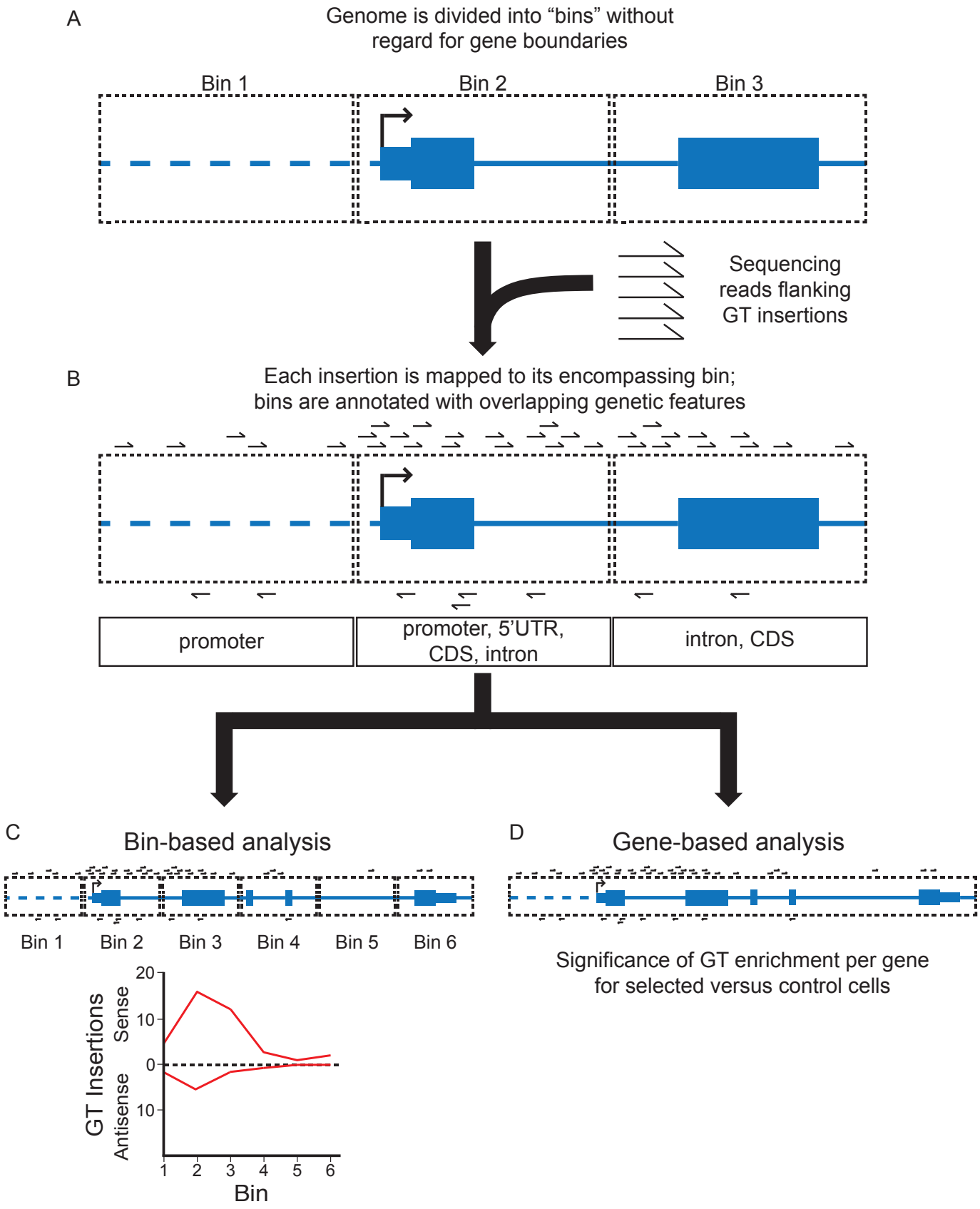
668 (A) A GT consists of direct long terminal repeats (LTRs), a strong splice acceptor (SA), a
669 reporter gene (mCherry) and a poly-adenylation (pA) sequence. A schematic of the 5' end of a
670 gene, including the promoter region, is also shown. (B) A GT can disrupt a gene by inserting into
671 an exon in the sense orientation (with respect to the coding sequence of the gene), interrupting
672 the coding sequence and causing premature transcriptional termination due to the pA sequence.
673 (C) An antisense GT insertion into an exon interrupts the coding sequence of the gene and
674 typically causes a frameshift mutation that leads to premature translational termination,
675 producing a truncated protein. (D) When a GT integrates into an intron in the sense orientation,
676 the SA causes the reporter gene and pA sequence to be spliced to the preceding exon, inevitably
677 leading to premature transcriptional termination due to the pA sequence. (E) An antisense GT
678 insertion in an intron will typically not disrupt a gene due to the directionality of the SA;
679 however, it could interfere with regulatory elements or with transcripts present on the antisense
680 strand. (F, G) GT insertions in the promoter region of a gene in either the sense or antisense
681 orientation generally do not affect transcription of the downstream gene. However, they could
682 potentially disrupt regulatory elements and alter transcription.

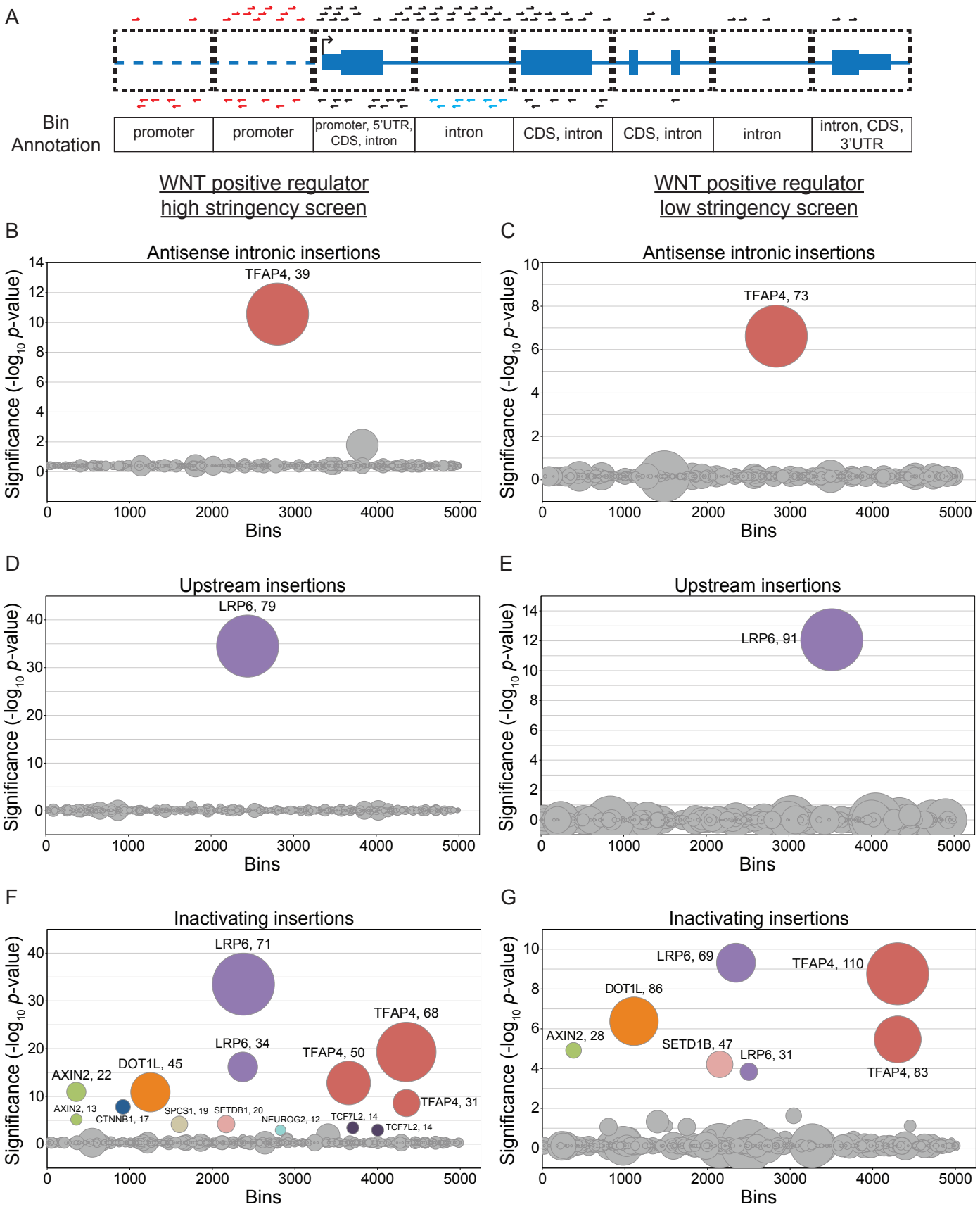
683 **S3 Fig. GT insertion patterns found in *TFAP4* and *DOT1L* in the WNT positive regulator**
684 **low stringency and high stringency screens.**
685 (A) The histogram indicates the number and orientation of insertions mapped to *TFAP4* in the
686 sorted cell populations from the WNT positive regulator low stringency and high stringency
687 screens. See legend to Fig 3A for details. (B) The histogram indicates the number and orientation
688 of insertions mapped to *DOT1L* (Chromosome 19, 2163750-2232749 bp) in unsorted cells and in
689 the sorted cell populations from the WNT positive regulator low stringency and high stringency
690 screens. The pattern of GT insertions seen in *DOT1L*, predominantly enriched for sense
691 insertions in the first intron, differs from the observed enrichment for both sense and antisense
692 insertions seen in the first intron of *TFAP4*.

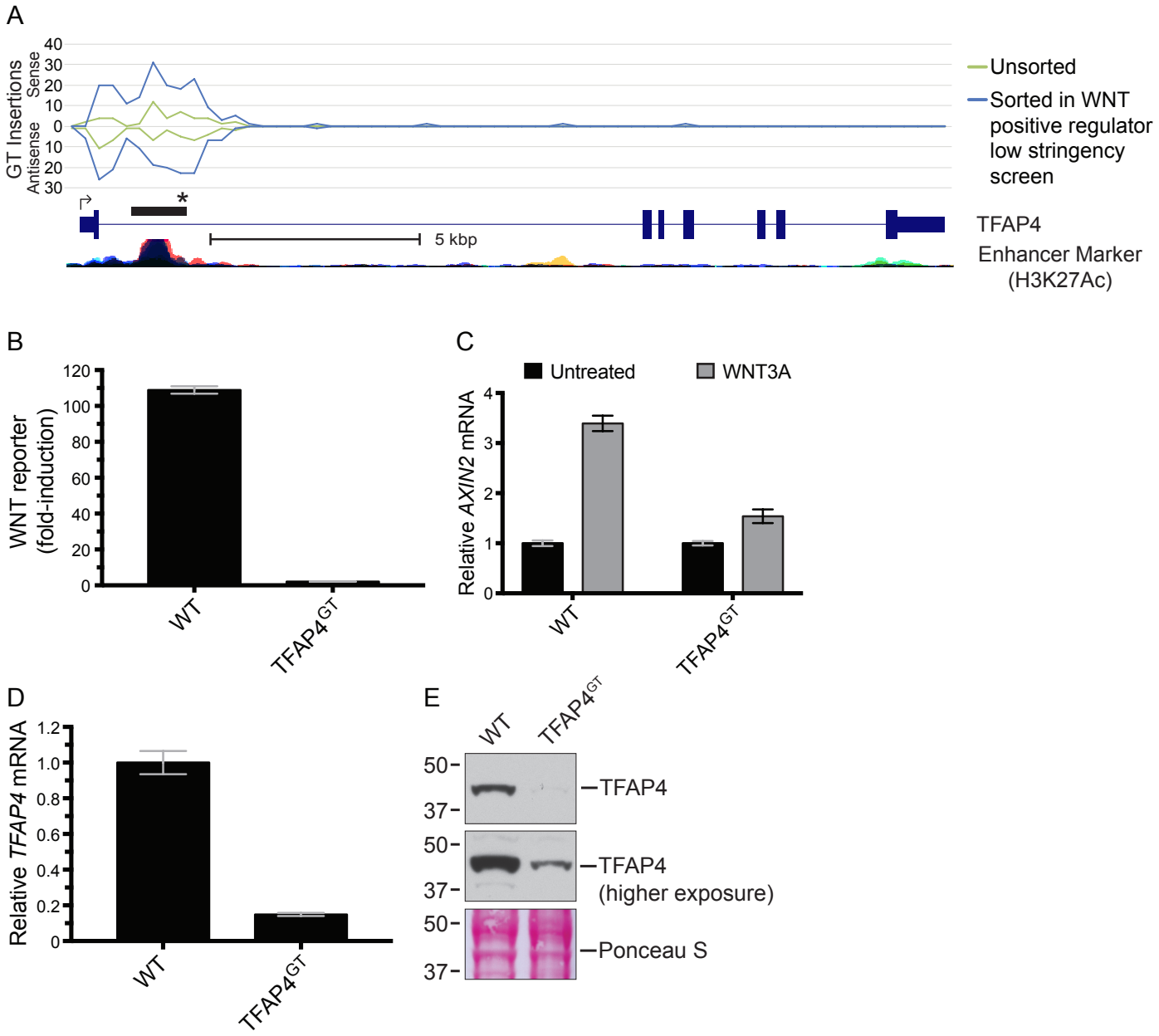
693 **S4 Fig. GT insertion patterns found in *LRP6* in the WNT positive regulator low stringency**
694 **and high stringency screens, and immunoblot analysis of LRP6.**
695 (A) The histogram indicates the number and orientation of insertions mapped to *LRP6* and to the
696 region ~12.5 kbp upstream of the TSS in the sorted cell populations from the WNT positive
697 regulator low stringency and high stringency screens. See legend to Fig 4A for details.
698 (B) The histogram shows an expanded view of the 5' end of *LRP6* and the region ~12.5 kbp
699 upstream of the TSS (left of the vertical dotted line), with traces for GT insertions mapped in the
700 sorted cell populations from the WNT positive regulator low stringency and high stringency
701 screens. See legend to Fig 4B for details. (C) Immunoblot analysis of LRP6. The top and bottom
702 parts of the same membrane were probed for LRP6 and ACTIN (loading control), respectively.
703 The cell lines from which the samples were prepared and loaded in duplicate are indicated above
704 the blots. Molecular weight standards in kDa are indicated to the left of each panel.

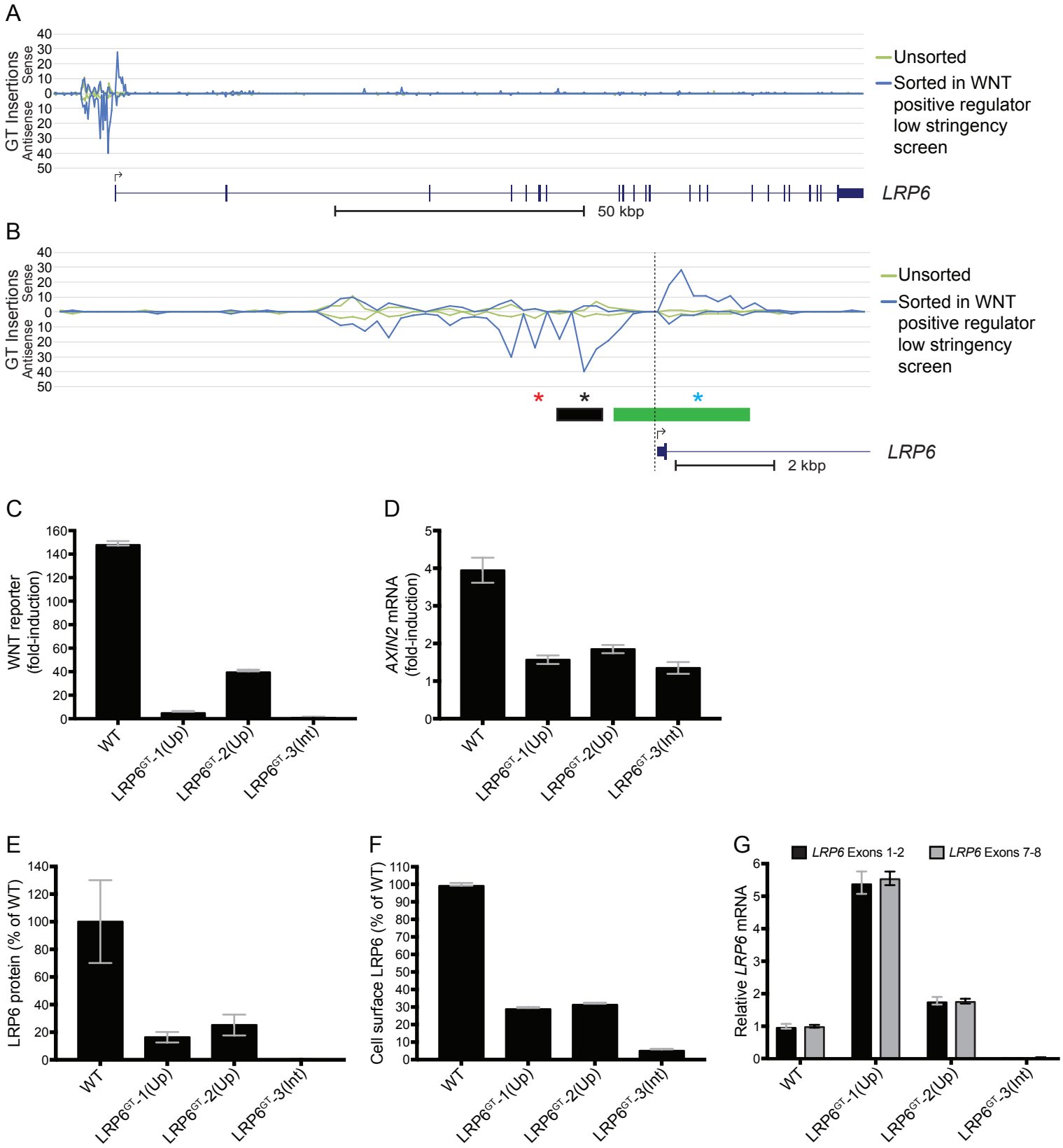
705 **S5 File. List of clonal cell lines containing GT insertions.**

706 The genomic sequences flanking GT insertion sites in the clonal cell lines used in this study are
707 described. The first column ("Clone Name") indicates the names of the clonal cell lines and the
708 second column ("Genomic sequence flanking GT") indicates the genomic sequences 5' and 3'
709 from the GT insertion site (relative to the sense orientation of the GT as described in S2 Fig).


A
Genome is divided into “bins” without regard for gene boundaries
Bin 1
Bin 2
Bin 3
Sequencing reads flanking GT insertions
B
Each insertion is mapped to its encompassing bin; bins are annotated with overlapping genetic features
promoter
promoter, 5’UTR, CDS, intron
intron, CDS
C
Bin-based analysis
Bin 1
Bin 2
Bin 3
Bin 4
Bin 5
Bin 6
GT Insertions
Sense
Antisense
20
10
0
10
1
2
3
4
5
6
Bin
D
Gene-based analysis
Significance of GT enrichment per gene for selected versus control cells


A
Bin Annotation
promoter
promoter
promoter, 5'UTR, CDS, intron
intron
CDS, intron
CDS, intron
intron
intron, CDS, 3'UTR
WNT positive regulator high stringency screen
WNT positive regulator low stringency screen
B
Antisense intronic insertions
TFAP4, 39
C
Antisense intronic insertions
TFAP4, 73
D
Upstream insertions
LRP6, 79
E
Upstream insertions
LRP6, 91
F
Inactivating insertions
LRP6, 71
TFAP4, 68
LRP6, 34
TFAP4, 50
DOT1L, 45
AXIN2, 22
AXIN2, 13
CTNNB1, 17
SPCS1, 19
SETDB1, 20
NEUROG2, 12
TCF7L2, 14
TCF7L2, 14
TFAP4, 31
G
Inactivating insertions
LRP6, 69
TFAP4, 110
DOT1L, 86
AXIN2, 28
SETD1B, 47
LRP6, 31
TFAP4, 83
Significance (-log10 p-value)
Bins


A
GT Insertions
Sense
Antisense
40
30
20
10
0
10
20
30
Unsorted
Sorted in WNT positive regulator low stringency screen
*
5 kbp
TFAP4
Enhancer Marker (H3K27Ac)
B
WNT reporter (fold-induction)
120
100
80
60
40
20
0
WT
TFAP4GT
C
Untreated
WNT3A
Relative AXIN2 mRNA
4
3
2
1
0
WT
TFAP4GT
D
Relative TFAP4 mRNA
1.2
1.0
0.8
0.6
0.4
0.2
0
WT
TFAP4GT
E
WT
TFAP4GT
50
37
TFAP4
50
37
TFAP4 (higher exposure)
50
37
Ponceau S


A
GT Insertions
Sense
Antisense
Unsorted
Sorted in WNT positive regulator low stringency screen
LRP6
50 kbp
B
GT Insertions
Sense
Antisense
Unsorted
Sorted in WNT positive regulator low stringency screen
LRP6
2 kbp
C
WNT reporter (fold-induction)
WT
LRP6GT-1(Up)
LRP6GT-2(Up)
LRP6GT-3(Int)
D
AXIN2 mRNA (fold-induction)
WT
LRP6GT-1(Up)
LRP6GT-2(Up)
LRP6GT-3(Int)
E
LRP6 protein (% of WT)
WT
LRP6GT-1(Up)
LRP6GT-2(Up)
LRP6GT-3(Int)
F
Cell surface LRP6 (% of WT)
WT
LRP6GT-1(Up)
LRP6GT-2(Up)
LRP6GT-3(Int)
G
LRP6 Exons 1-2
LRP6 Exons 7-8
Relative LRP6 mRNA
WT
LRP6GT-1(Up)
LRP6GT-2(Up)
LRP6GT-3(Int)