RADAR: Annotation and prioritization of variants in the post-transcriptional regulome for RNA-binding proteins

Jing Zhang[1,2*], Jason Liu[1,2*], Donghoon Lee[1], Lucas Lochovsky[1], Jo-Jo Feng[3], Shaoke Lou[1,2], Michael Schoenberg[1,2,3], Mark Gerstein[1,2,3]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA
[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA
[3]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA

RCR20

Style Definition: Normal: Font:11 pt, Font color: Text 1, Justified, Line spacing: double

Style Definition: Heading 1: Line spacing: double

Style Definition: Heading 2: Font:Times New Roman, 14 pt, Bold, Font color: Auto, Justified, Space After: 6 pt, Line spacing: double

Style Definition: Heading 3: Font:(Default) Times New Roman, Bold, Font color: Auto, Justified, Line spacing: double

Style Definition: Comment Text: Font:11 pt, Font color: Text 1, Justified, Line spacing: double

Deleted: An integrative framework for variant annotation

Deleted:

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

Abstract

##############

JZ2MG: please compare to the editor's version, which I don't like. First sentence is important, and I don't want to waste on introducing other ppl's work.

###############

Efforts to annotate and pinpoint deleterious variants have advanced our understanding of transcriptional processing. However, the impact of variants in the post-transcriptional regulome has been understudied. RNA binding proteins (RBPs) play key roles in post-transcriptional regulation, and their binding sites cover a larger amount of the genome than coding exons. Thus, we aimed to construct a comprehensive RBP regulome by integrating the full catalog of eCLIP, RNA Bind-n-Seq, and shRNA RNA-Seq from ENCODE. In addition, we developed a variant impact scoring framework, RADAR (***R***NA Bin***D***ing protein regulome ***A***nnotation and p***R***ioritization), that incorporates human and cross-species conservation, RNA structure, network centrality, and RBP motifs, as well as user input, such as differential expression profiles or prior knowledge of regulators. Our results demonstrate that RADAR can successfully pinpoint intronic, splicing-disruptive variants in key genes such as TP53 that cannot be detected by current methods.

################ OLD version which I like most ################

RNA Binding proteins (RBP) play key roles in post-transcriptional regulation. Their binding sites cover a larger amount of the genome than coding exons, but most of the current variant prioritization methods ignore RBP regulatory effects and only focus on transcriptional-level regulation. Here, we integrated the full catalog of eCLIP, RNA Bind-n-Seq, and shRNA RNA-Seq experiments from ENCODE to deeply annotate the RBP regulome. We first proposed a variant impact scoring framework, RADAR, which uses human and cross-species conservation, RNA structure, network centrality and RBP motifs to provide a baseline impact score. Then RADAR incorporated user-specific inputs, such as differential expression/mutational profiles or prior knowledge of regulators or genes to further highlight

FRAC

BINDING

Formatted: Font:26 pt

Formatted: Heading 1, Left, None, Space Before: 0 pt, After: 0 pt

Formatted: Font:26 pt, Not Bold

Deleted: ... [1]

Formatted: Font:11 pt

Formatted: Font:11 pt

Deleted: RBP) can cause numerous diseases

Formatted: Left

Deleted: the effect of variants in their regulome has been barely investigated. Hence

Deleted: 318

Deleted: 76

Deleted: 472

Deleted: the new release of the

Deleted: project

Deleted: First, we showed that around 90 percent of RBPs are enriched with rare variants in their binding sites, suggesting extensive purifying selections.

Deleted: then

Deleted: by combining RBP binding,

Deleted: context,

Deleted: , and conservation features with polymorphism data

Deleted: we

Deleted: patient survival,

Deleted: ,

Deleted: and

Deleted: to reweight the variants

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

disease- and tissue-specific variants. Results on somatic and germline variants demonstrate that RADAR can successfully pinpoint intronic, splicing-disruptive variants in key genes such as TP53, which cannot be fully detected by current methods

Deleted: causal ones

Deleted: both

Deleted: and somatic variant datasets

Deleted: disease-relevant variants and uncover the underlying regulation mechanism of post-transcriptional regulation.

Formatted: Font:11 pt

Deleted:

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

Introduction

Dysregulation of gene expression is a hallmark of many diseases, including cancer[1]. In recent years, the accumulation of transcription-level functional characterization data, such as transcriptional factor binding, chromatin accessibility, histone modification, and methylation, has brought great success to annotating and pinpointing deleterious variants. However, beyond transcriptional processing genes also experience a series of delicately controlled steps, including the conversion of premature RNA to mature RNA, and then the transportation, translation, and degradation of RNA in the cell. Dysregulation in any one of these steps can alter the final fate of gene products and result in abnormal phenotypes[2-4]. Furthermore, this post-transcriptional regulome covers an even larger amount of the genome than coding exons and demonstrates significantly higher cross-population and cross-species conservation. Unfortunately, the impact of variants in the post-transcriptional regulome has been barely investigated, partially due to the lack of large-scale functional mapping.

RNA binding proteins (RBPs) have been reported to play essential roles in both co- and post-transcriptional regulation[5-7]. RBPs bind to thousands of genes in the cell through multiple processes, including splicing, cleavage and polyadenylation, editing, localization, stability, and translation[8-12]. Recently, scientists have made efforts to complete these post- or co-transcriptional regulomes by synthesizing public RBP binding profiles[13-16], which have greatly expanded our understanding of RBP regulation. Since 2016, the Encyclopedia of DNA Elements (ENCODE) consortium began mapping the post-transcriptional regulome in 2016 using various types of assays on matched cell types. ENCODE has released large-scale enhanced crosslinking and immunoprecipitation (eCLIP) experiments for hundreds of RBPs[17]. This methodology provides high-quality RBP binding profiles with strict quality control and uniform peak calling to accurately catalog the RBP binding sites at a single nucleotide resolution. Simultaneously, ENCODE performed

Deleted: and Background
Formatted ... [6]
Formatted ... [5]
Formatted ... [7]
Formatted ... [8]
Deleted:
Formatted ... [9]
Deleted: after (or simultaneously while) DNA has be ... [10]
Formatted ... [11]
Deleted: processing, such as
Formatted ... [12]
Deleted: followed by
Formatted ... [13]
Deleted: then
Formatted ... [14]
Deleted: of
Formatted ... [15]
Deleted: may
Formatted ... [16]
Deleted: -6.… Despite its importance ... [17]
Formatted ... [18]
Deleted: regulation, such
Formatted ... [19]
Deleted: underdeveloped
Formatted ... [20]
Deleted: its less systematic
Formatted ... [21]
Deleted: during
Formatted ... [23]
Formatted ... [22]
Deleted: regulations
Formatted ... [24]
Deleted: -9.… They ... [25]
Formatted ... [26]
Deleted: RNA
Formatted ... [27]
Deleted: 10-14…-12. Recently, many efforts ... [28]
Formatted ... [29]
Deleted: been
Formatted ... [30]
Deleted: 15-18
Formatted ... [31]
Deleted: started to map
Formatted ... [32]
Deleted: First,
Formatted ... [33]
Deleted: CLIP
Formatted ... [34]
Deleted: 19…7. It ... [35]
Formatted ... [36]
Deleted: It also simultaneously
Formatted ... [37]
Formatted ... [2]
Formatted ... [3]

expression quantification by RNA-Seq after knocking down various RBPs. Finally, ENCODE has quantitatively assessed the context and structural binding specificity of many RBPs by Bind-n-Seq experiments.[18]

CONS

In this study, we aimed to construct a comprehensive RBP regulome and a scoring framework to annotate and prioritize variants within it. We collected a full catalog of 318 eCLIP (for 112 RBPs), 76 Bind-n-Seq, and 472 shRNA RNA-Seq experiments from ENCODE to construct a comprehensive post-transcriptional regulome. By combining polymorphism data from large sequencing cohorts, like the 1,000 Genomes Project, we demonstrated that 88 and 94 percent of RBPs showed an enrichment of rare variants in coding and noncoding regions, respectively, compared to the genomic average. This strongly indicates the purifying selection of the RBP regulome. Furthermore, we developed a top-down scoring scheme, named RADAR (***R***N***A*** Bin***D***ing Protein regulome ***A***nnotation and p***R***ioritization), to investigate the variant impact in such regions. RADAR first combines RBP binding, conservation, network, and motif disruption features with polymorphism data to quantify variant impact described by a universal baseline score. Then, it allows tissue- or disease-specific inputs, such as differential expression, somatic mutation, and prior knowledge of genes, to further highlight relevant variants (Fig. 1). By applying RADAR to both somatic and germline variants from disease genomes, we demonstrate that it can pinpoint disease-associated variants missed by other methods. Thus, RADAR provides an effective approach to analyzing genetic variants in the RBP regulome, and can be leveraged to expand our understanding of post-transcriptional regulation. To this end, we have implemented the RADAR annotation and prioritization scheme into software for community use (radar.gersteinlab.org).

Results

Deleted: of

Formatted: Font:11 pt

Deleted: Also

Deleted: also performed a quantitative assessment of

Formatted: Font:11 pt

Formatted: Font:11 pt

Formatted: Font:11 pt

Deleted: \cite{24837674}.

Formatted: Font:11 pt

Deleted: JZ2MG: whenever I talked about conservati ... [39]

Formatted: Font:11 pt

Deleted: paper

Formatted ... [40]

Deleted: the

Formatted ... [41]

Deleted: ,

Formatted ... [42]

Deleted: 1000

Formatted: Font:11 pt

Deleted: noticeably higher conservations

Formatted: Font:11 pt, Font color: Text 1

Deleted: .

Formatted: Font:11 pt

Deleted: proposed

Formatted ... [43]

Deleted: structure, context

Formatted: Font:11 pt

Deleted: conservation

Formatted ... [44]

Deleted: Figure

Formatted: Font:11 pt

Deleted: our scoring scheme on

Formatted ... [45]

Deleted: RADAR is able to

Formatted: Font:11 pt

Deleted: Finally

Formatted ... [46]

Deleted: 2

Formatted ... [47]

Formatted ... [48]

Formatted ... [38]

Formatted: Right: 0"

Defining the RBP regulome using eCLIP data

We used the binding profiles of 112 distinct RBPs from ENCODE to fully explore the human RBP regulome (Supplementary Table 1), which has been previously underinvestigated. Many of these RBPs are known to play key roles in post-transcriptional regulation, including splicing, RNA localization, transportation, decay, and translation (Supplementary Fig. 1).

Our definition of the RBP regulome covers 52.6 Mbp of the human genome after duplicate and blacklist removal (Fig. 2A). This is 1.5 and 5.9 times the size the whole exome and lincRNAs, respectively. In addition, only 53.1% of the RBP regulome has transcription-level annotations, such as transcription binding sites, open chromatin regions, and enhancers (Supplementary Fig. 2). Unlike the transcription regulome, which has many distal elements, 55.1% of the RBP regulome is in the immediate neighborhood of the exome regions, such as coding exons, 3' or 5' untranslated regions (UTRs), and nearby introns (Fig. 2B; see methods section for more details). Furthermore, we observed significantly higher conservation scores in the peak regions versus the non-peak regions in almost all annotation categories, providing additional evidence of regulatory roles of RBP peaks (Fig. 2C). In summary, the large size of the regulome, the limited overlap with previous annotations, and the elevated conservation scores highlight the necessity of our computational efforts to define the RBP regulome.

Using universal features for baseline RADAR score

To annotate and prioritize variants in RBP binding sites, we built a baseline score framework for RADAR that includes three components: (1) sequence and structure conservation; (2) network centrality; and (3) nucleotide impact from motif analysis.

Deleted: through
Formatted ... [51]
Formatted ... [52]
Deleted: Here we first
Formatted ... [53]
Formatted ... [54]
Deleted: Table S1).
Formatted ... [55]
Deleted: and
Formatted ... [56]
Deleted: S1
Formatted ... [57]
Comment [j1]: JZ2JL please check this number
Deleted: While under-investigated, our defined
Formatted ... [58]
Deleted: a decent part
Formatted ... [59]
Deleted: (52.6Mbp
Formatted ... [60]
Deleted: ,
Formatted ... [61]
Deleted: It
Formatted ... [62]
Deleted: of
Formatted ... [63]
Deleted: Besides
Formatted ... [64]
Deleted: percent
Formatted ... [65]
Deleted: is overlapped by the
Formatted ... [66]
Deleted: including
Formatted ... [67]
Deleted: S2
Formatted ... [68]
Deleted: percent
Formatted ... [69]
Deleted: located
Formatted ... [70]
Deleted: ,
Formatted ... [71]
Deleted: , details
Formatted ... [72]
Deleted:).
Formatted ... [73]
Deleted: in almost all annotation categories,
Formatted ... [74]
Deleted: PhastCons
Deleted: vs.
Formatted ... [75]
Deleted: ,
Formatted ... [76]
Deleted: their
Formatted ... [77]
Formatted ... [78]
Formatted ... [79]
Formatted ... [80]
Formatted ... [81]
Formatted ... [82]
Formatted ... [83]
Formatted ... [84]
Formatted ... [85]
... [86]

REF

Sequence and structure conservation in the RBP regulome

Cross-species sequence comparisons have been widely used to discover regions with biological functions. For example, GERP score maps the human genome to other species to identify nucleotide-level evolutional constraints[19,20]. We used the GERP score in our baseline RADAR framework to detect potentially deleterious mutations in the RBP regulome.

Because the enrichment of rare variants indicates a purifying selection in functional regions in human genomes[21-23], we also inferred conservation of RBP binding sites by integrating population-level polymorphism data from large cohorts (i.e. the 1,000 Genomes Project)[24,25]. GC percentage may confound such inference by introducing read coverage variations, which is a sensitive parameter in the downstream variant calling process[26,27]. Therefore, we calculated the fraction of rare variants, defined as those with derived allele frequencies (DAFs) less than 0.5%, within each RBP binding site and compared them with those from regions with similar GC content as the background (see methods section for more details). In total, 88.4% of the RBPs (99 out of 112) showed elevated rare variant fraction in coding regions after GC correction (Fig. 3A). Similarly, in the noncoding part of the binding sites, 93.8% of RBPs (105 out of 112) exhibited an enrichment of rare variants. This observation convincingly demonstrates the accuracy of our RBP regulome definition (Supplementary Table 2).

Some well-known disease-causing RBPs demonstrate the largest enrichment of rare variants. For example, the oncogene XRN2, which binds to the 3' end of transcripts to degrade aberrantly transcribed isoforms, showed significant enrichment of rare variants in its binding sites[28]. Specifically, it demonstrates 12.7% and 10.3% more rare variants in coding and noncoding regions, respectively (adjusted P values are 1.89×10^{-9} and 2.85×10^{-118} for one-sided binomial tests)[29]. Hence, we used the enrichment of rare variants to infer the

Formatted ... [91]
Deleted: [34-36…1-23]. Hence, aside from the cross-specie ... [92]
Deleted: the purifying selection on the
Formatted ... [93]
Formatted ... [94]
Deleted: ,
Formatted ... [95]
Deleted: g.
Formatted ... [96]
Deleted: 1000
Formatted ... [97]
Deleted: cite {26432245, 23128226}.
Formatted ... [98]
Deleted: because it causes
Formatted ... [99]
Deleted: [37,38…6,27]. Hence ... [100]
Formatted ... [101]
Deleted: first
Formatted ... [102]
Deleted: (
Formatted ... [103]
Deleted: frequency (DAF
Formatted ... [104]
Deleted: %)
Formatted ... [105]
Deleted: RBP's
Formatted ... [106]
Deleted: sites,
Formatted ... [107]
Deleted: it
Formatted ... [108]
Deleted: a
Formatted ... [109]
Deleted: in methods
Formatted ... [110]
Deleted: percent
Formatted ... [111]
Deleted: show
Formatted ... [112]
Deleted: compared to those of the background regions
Formatted ... [113]
Deleted: percent
Formatted ... [114]
Deleted: exhibit
Formatted ... [115]
Deleted: . (
Formatted ... [116]
Deleted: S2
Formatted ... [117]
Deleted: RNA binding proteins
Formatted ... [118]
Deleted: rare DAF
Formatted ... [119]
Deleted: [39]
Formatted ... [120]
Formatted ... [121]
... [122]
Formatted ... [123]
Formatted ... [88]
Formatted ... [89]

selection pressure in RBP binding sites to weight the variants in such regulator regions (see methods for more details).

RNA secondary structures have been reported to affect every step of protein expression and RNA stability[30]. We incorporated structural features predicted by Evofold, which uses phylogenetic stochastic context-free grammars to identify functional RNAs in the human genome that are deeply conserved across species[31]. We found that the RBP binding sites demonstrated significantly higher conversation after intersecting with conserved structural regions defined by Evofold. Thus, we used the Evofold regions in our baseline score.

Incorporating network information

Highlighting variants in binding hubs

It has been reported that genes within network hubs demonstrate enrichment of rare variants—a sign of strong purifying selection[21,22,32]. We hypothesized that RBP binding hubs would show similar characteristics because once mutated larger regulation alterations may be introduced. To test this, we separated the regulome based on the number of associated RBPs. Most regulome regions (62%) were associated with only one RBP (Fig. 3B and Supplementary Fig. 4). As the number of RBPs increased, we observed a clear trend of larger rare variant enrichment. For instance, noncoding regions with at least five or 10 RBPs exhibited 2.2% or 13.4% more rare variants, respectively (top 5% and 1%, Fig 3C). This observation supports our hypothesis that the RNA regulome hubs are under stronger selection pressure and thus, should be given high priority when evaluating the functional impacts of mutations.

Up-weighting genes differentially expressed after RBP knockdown

RNA-seq expression profiling before and after shRNA mediated RBP depletion from ENCODE can help to infer the gene expression changes introduced by RBP knockdown. Variants with disruptive effects on

Formatted ... [126]
Moved (insertion) [1] ... [127]
Deleted: 2.2.2 Highlight
Formatted ... [129]
Deleted: RBP
Formatted ... [128]
Formatted ... [130]
Deleted: usually exhibit larger
Formatted ... [131]
Formatted ... [132]
Formatted ... [133]
Deleted: $^{34,35,44\ldots1,22,32}$. Similarly, we suspect ... [134]
Formatted ... [135]
Deleted: may demonstrate
Formatted ... [136]
Deleted: hypothesis
Formatted ... [137]
Deleted: The majority of the
Formatted ... [138]
Deleted: percent) are
Formatted ... [139]
Deleted: 1
Formatted ... [140]
Deleted: S4
Formatted ... [141]
Deleted: in the
Formatted ... [142]
Deleted: , regions
Formatted ... [143]
Deleted: 5 and
Formatted ... [144]
Deleted: and
Formatted ... [145]
Deleted: percent
Formatted ... [146]
Deleted: percent,
Formatted ... [147]
Deleted: convincingly
Formatted ... [148]
Deleted: ,
Formatted ... [149]
Deleted: hence
Formatted ... [150]
Deleted: 2.2.3 Motif analysis for nucleotide impact ... [151]
Moved down [2]: RADAR consists of two sources ... [152]
Formatted ... [153]
Deleted: First, it has been reported that many of the ... [154]
Formatted ... [155]
Moved up [1]: ... [156]
Formatted ... [157]
Deleted: the protein expression and RNA stability ... [158]
Formatted ... [159]
Deleted: genes
Formatted ... [160]
Formatted ... [161]
Formatted ... [162]
Formatted ... [124]
Formatted ... [125]

RBP binding may affect or even completely remove the RBP binding and hence affect gene expressions in a similar way. Therefore, we extracted the differentially expressed genes from RNA-Seq before and after shRNA-mediated RBP depletion. Then, we up-weighted all variants that were located near differentially expressed genes and simultaneously disrupted the binding of the corresponding RBPs (schematic in Supplementary Fig. 5).

Using motif analysis to determine nucleotide impact

Mutations that change the RBP binding affinity may alter RBP regulation via motif disruption. We quantified the difference of position weight matrix scores of the mutant allele against the reference allele. RADAR consists of two sources of motifs. First, we used the motifs identified from RNA Bind-n-Seq experiments from ENCODE because it has been reported that many RBP binding events *in vivo* can be captured by binding preferences *in vitro*. Second, we used the *de novo* motifs discovered directly from binding peaks using the default settings in DREME (details see methods). For each variant, we quantified the nucleotide effect using the highest motif score from these two sources.

Incorporating user-specific features to reweight variant impact

Variant Prioritization can be improved if informative priors can be appropriately incorporated in to the scoring system. Therefore, our RADAR framework allows various types of user-input to help identify disease-relevant variants. Specifically, we adopted a top-down scheme to incorporate regulator, element, and variant level information to up-weight any factors that are possibly associated with target disease.

Highlighting key regulators through expression profiles

Key regulators are often associated with disease progression, so variants that affect such regulation should be prioritized. Therefore, we used RADAR to find disease-specific expression profiles, RADAR tried to find such key regulators by combining the RBP regulatory network information with expressions.

Deleted: such experiments
Formatted: Font:11 pt
Deleted: are
Formatted: Font:11 pt
Deleted: weight
Formatted: Font:11 pt
Deleted: such
Formatted: Font:11 pt
Formatted: Font:11 pt
Deleted: at the same time disrupt
Formatted: Font:11 pt
Formatted: Font:11 pt
Deleted: S5
Formatted: Font:11 pt
Formatted: Font:11 pt
Moved (insertion) [2]
Formatted: Font:11 pt
Deleted: 2.3 User

Formatted: Heading 2, Left, None, Space Before: 0 pt, After: 0 pt
Formatted: Font:14 pt, Font color: Auto
Deleted: 2.3.1 Use

Formatted: Heading 3, Left
Formatted: Font:12 pt, Font color: Auto
Deleted: **to prioritize key regulators**
Formatted: Font:12 pt, Font color: Auto
Deleted: In many diseases, several key
Formatted: Font:11 pt
Deleted: can be
Deleted: and
Formatted: Font:11 pt
Formatted: Font:11 pt
Deleted: given additional
Formatted: Font:11 pt
Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around
Formatted: Right: 0"

Specifically, we first constructed the RBP network from the eCLIP profiles and defined differentially expressed genes from disease and normal cell types. Then for each RBP, we quantified the regulation power for each RBP by association with aggregated disease-to-normal differential expressions from many samples. We applied this approach on 19 cancer types from TCGA and the regulation powers are given in Fig. 4. We found that among many of the RBPs with larger for each RBP power have been reported as cancer-associated genes (with * in Fig. 4B, Table S3). Interestingly, the regulatory power of two key genes PPIL4 and SUB1 were found to be significantly associated with patient survival (Fig. 4C).

In our RADAR framework, we further highlight variants that are associated with significant regulators in their corresponding cancer types by adding extra point to their baseline scores (details see methods). We can easily extend such analysis for other diseases by incorporating differential expression profiles from others cohorts such as GTEx[33,34].

Up-weighting key elements from either prior knowledge or mutational profiles

RADAR reconsiders the functional impact difference among RBP peaks either by their associated genes or cohort-level mutational profiles. Users can input a prioritized gene list, such as well-documented risk genes for a disease of interest, and RADAR up-weights all the RBP peaks that are close to these genes. Genes that undergo significant expression or epigenetic changes are mostly cell type specific and be can be used to highlight more relevant variants.

GRAMA

In addition, RADAR can incorporate variant recurrence, which has been widely used to discover key disease regions, to reweight different RBP peaks. Regions with more mutations than expected are often considered to be disease-driving[35-37]. For example, we used RADAR to define a local background somatic

MENT

Deleted: differential expressions
Formatted ... [165]
Deleted: its
Formatted ... [166]
Deleted: sampls
Formatted ... [167]
Formatted ... [168]
Formatted ... [169]
Deleted: .
Formatted ... [170]
Deleted: normal
Formatted ... [171]
Deleted: \cite{JZ2JL: GTEx nature marker papers}.
Deleted: 2.3.2 Prior knowledge of regulators, target ... [172]
Formatted ... [173]
Deleted: of the regulators, target genes, and varia ... [174]
Formatted ... [175]
Deleted: splicing factor QKI has been reported by m ... [176]
Formatted ... [177]
Deleted: known to play critical roles in cancer. Cell ... [178]
Formatted ... [179]
Deleted: -
Formatted ... [180]
Deleted: -
Formatted ... [181]
Moved (insertion) [3] ... [182]
Formatted ... [183]
Deleted: Hence, our RADAR framework allows use ... [186]
Formatted ... [184]
Formatted ... [187]
Formatted ... [185]
Deleted: in disease
Formatted ... [188]
Deleted: mutations
Formatted ... [189]
Deleted: as
Formatted ... [190]
Deleted:
Formatted ... [191]
Deleted: events \cite{mutsigCV, larva, moat}. Take ... [192]
Formatted ... [193]
Deleted: . Given the
Formatted ... [194]
Formatted ... [163]
Formatted ... [164]

mutation rate from a large cohort of cancer patients to evaluate the mutation burden in each RBP peak. Variants that are associated with burdened elements are given higher priority in the scoring scheme.

Prioritizing variants with a RADAR weighted scoring scheme

By integrating the universal and user-specific information described above, our entropy-based scoring scheme investigates the functional impacts of variants that are specific to post-transcriptional regulation (Fig. 1, Table 1, and Supplementary Fig. 5). First, RADAR adds up the (weighted) score of variants for all universal features, which include sequence and structure conservation, binding hubs, and motif disruptions. Then, depending on the user inputs, RADAR further up-weights variants with mutations in the key RBP binding sites, nearby genes of interest, or within regulatory elements.

Table 1. Features used by RADAR

Category	Feature	Source	Scoring Scheme
Universal	Selection pressure	eCLIP	Weighted-entropy
	Binding hotspots	eCLIP	Weighted-entropy
	RBP-gene association	shRNA RNA-seq	Entropy
	Motif disruption	Bind-n-Seq DREME	Weighted-entropy
	Structure sensitivity	Evofold	Entropy
	Conservation	Gerp	Entropy
User-specific	RBP regulatory power	Survival Expression	Entropy
	Key genes	Prior knowledge	Entropy
	Mutation Recurrence	Mutation profiles	Entropy

Applying RADAR to pathological germline variants

We calculated baseline RADAR scores on all pathological variants from HGMD. We used the 1,000 Genomes variants as the background to compare the distribution of scores. As expected, the HGMD variants scored significantly higher than somatic mutations (Supplementary Fig. 6). For example, the mean RADAR

Formatted ... [197]
Deleted: peaks
Formatted ... [199]
Formatted ... [200]
Deleted: prioritization
Formatted ... [201]
Deleted: 2.4
Deleted: to prioritize variants
Formatted ... [202]
Formatted ... [203]
Deleted: mentioned
Formatted ... [206]
Deleted: we proposed an
Formatted ... [204]
Formatted ... [207]
Formatted ... [205]
Deleted: to investigate
Formatted ... [208]
Deleted: S5
Formatted ... [209]
Deleted: we added
Formatted ... [210]
Deleted: entropy
Formatted ... [211]
Deleted: rare variant enrichment
Formatted ... [212]
Deleted: hotspot, structure,
Deleted: , and conservations.
Formatted ... [213]
Formatted ... [214]
Deleted: we
Formatted ... [215]
Deleted: weight
Formatted ... [216]
Deleted: that fall into
Formatted ... [217]
Deleted: with more than expected variants
Formatted ... [218]
Formatted ... [219]
Deleted:
Formatted ... [220]
Formatted ... [221]
Formatted Table ... [222]
Formatted ... [225]
Formatted ... [226]
Formatted ... [227]
Formatted ... [223]
Formatted ... [224]
Formatted ... [228]
Formatted ... [229]
Formatted ... [230]
Formatted ... [233]
Formatted ... [234]
Formatted ... [231]
Formatted ... [232]
Formatted ... [235]
Formatted ... [236]
Formatted ... [237]
Formatted ... [238]
Formatted ... [239]
Formatted ... [240]
Formatted ... [241]
Formatted ... [242]
Formatted ... [243]
Formatted ... [244]

score for HGMD variants is 0.445, while it is only 0.044 for 1,000 genomes variants (P value <2.2e-16 for two sided Wilcoxon test). We further compared RADAR scores of HGMD variants with other methods (Supplementary Table 6). In total, 992 HGMD variants were identified by only our methods, 29.6% of which were noncoding variants located in a nearby intron, 5'UTR, or 3'UTR (and their extended regions). An example of such a variant is shown in Fig. 5; this variant is located 28 base pairs away from the acceptor site of exon 3 in TP53. eCLIP experiments showed strong binding evidence in seven RBPs, most of which are splicing factors. The co-binding of these above-mentioned splicing factors strongly indicate that this is a key splicing regulatory site. Specifically, the A to T mutation strongly disrupts the binding motif of SF3B4, increasing the possibility of splicing alteration effects. Our finding is not reflected in previous methods for variant prioritization.

Applying RADAR to somatic variants in cancer

We next aimed to leverage our scheme to evaluate the deleteriousness of somatic variants from public datasets. Due to the lack of a gold standard, we evaluated our results from two perspectives. First, we reasoned that since hundreds of cancer-associated genes are known to play essential roles through various pathways[38,39], variants associated with these genes are likely to have the highest functional impact[21]. To test this hypothesis, we first linked each variant with a gene by the shortest distance according to the Gencode v19 annotation. We tested four cancer types, breast, liver, lung, and prostate cancer, and found in all cases that variants associated with cancer-related genes showed significant enrichment, with a larger RNA level functional impact (Supplementary Fig. 7). For example, we found a 3.27- and 3.36-fold increase in high-impact variants at a threshold level of 2.5 and 3, respectively, in breast cancer patients (P < 2.2e-16, single-sided Wilcoxon).

Formatted ... [248]
Deleted: S6). Specifically, we found
Deleted: that are explained
Deleted: and
Formatted ... [249]
Formatted ... [250]
Formatted ... [251]
Deleted: them are
Formatted ... [252]
Deleted: that are
Deleted: the
Formatted ... [254]
Deleted:
Formatted ... [255]
Deleted: and
Formatted ... [253]
Formatted ... [256]
Deleted: variants
Formatted ... [257]
Deleted: given
Formatted ... [258]
Deleted: . It
Formatted ... [259]
Deleted: bp
Formatted ... [260]
Deleted:
Formatted ... [261]
Deleted: 7
Formatted ... [262]
Deleted:
Formatted ... [263]
Deleted: this
Formatted ... [264]
Deleted: 2.6 Application on
Formatted ... [265]
Formatted ... [266]
Deleted: 2.6.1 Somatic variants associated with CO ... [267]
Formatted ... [268]
Deleted: applied
Formatted ... [269]
Deleted: golden
Formatted ... [270]
Deleted: evaluate
Formatted ... [271]
Deleted: aspects
Formatted ... [272]
Deleted: gene
Formatted ... [273]
Deleted: [47,48…8,39]. Hence, in general, ... [274]
Formatted ... [275]
Deleted: supposed
Formatted ... [276]
Deleted: a higher
Formatted ... [277]
Formatted ... [278]
Formatted ... [279]
Formatted ... [280]
Formatted ... [281]
Formatted ... [282]
Formatted ... [283]
Formatted ... [284]
Formatted ... [285]

In our second approach, we hypothesized that variant recurrence could be a sign of functionality and may indicate an association with cancer[21-23]. Thus, we compared the variants' score distribution from RNA binding peaks with or without recurrence. Specifically, we separated the peaks of variants from more than one sample from those that were mutated in only one sample, and then compared the percentage of high-impact scores. We found that in most cancer types, elements with recurrent variants were associated with a larger fraction of high-impact mutations. For example, in breast cancer recurrent elements demonstrated a 1.20, 1.55, and 1.77-fold enrichment of high-impact variants with RADAR greater than 1.5, 2.5, and 3.0, respectively, resulting in a P value of 1.71e-19 (one-sided Wilcoxon test).

A case study on breast cancer patients

We applied our method on a set of breast cancer somatic variants from 963 patients released by Alexandrov *et al*[40]. We used COSMIC genes and expression and mutational profiles as additional features. In total, we determined that around 3% of the 68,000 variants alter post-transcriptional regulation to some degree. Specifically, 169 out of the 501 highly ranked variants only reported by our tool were located in noncoding regions, with 15, 28, and 24 from nearby introns, the 5' UTR, and the 3' UTR, respectively (Supplementary Fig. 8). We found that variants in the intronic region usually bind within 30 bp of the splice sites and break the motifs of many splicing factor binding sites. For the 3' UTR regions, variants reported only by RADAR were within the binding peaks of cleavage stimulation factor binding sites, strongly indicative of a role in the polyadenylation of pre-mRNAs. The discovery of such meaningful results indicates the ability of RADAR to differentiate deleterious mutations that disrupt post-transcriptional regulation.

Discussion

In this study, we integrated the full catalog of eCLIP, Bind-n-Seq, and shRNA RNA-Seq experiments from ENCODE to build a RNA regulome for post-transcriptional regulation. Although DNA-level regulation takes

Formatted ... [291]
Deleted: is considered
Formatted ... [290]
Formatted ... [293]
Formatted ... [294]
Deleted: [34-36…1-23], ... [295]
Formatted ... [296]
Deleted: also
Formatted ... [297]
Deleted: with
Formatted ... [298]
Deleted: are
Formatted ... [299]
Deleted: higher
Formatted ... [300]
Deleted: are
Formatted ... [301]
Deleted:
Formatted ... [302]
Deleted: Breast
Formatted ... [303]
Deleted: ,
Formatted ... [304]
Deleted: factor of
Formatted ... [305]
Deleted:
Formatted ... [306]
Deleted: at
Formatted ... [307]
Deleted: from
Formatted ... [308]
Deleted: .
Deleted: 2.6.2
Formatted ... [309]
Formatted ... [310]
Formatted ... [311]
Formatted ... [312]
Deleted: [50] and
Formatted ... [313]
Deleted: ,
Formatted ... [314]
Deleted:
Formatted ... [315]
Deleted: percent out
Formatted ... [316]
Deleted: 68k
Formatted ... [317]
Deleted: was evaluated to
Formatted ... [318]
Deleted: regulations
Formatted ... [319]
Deleted: are
Formatted ... [320]
Deleted: the
Formatted ... [321]
Formatted ... [322]
Formatted ... [323]
Formatted ... [324]
Formatted ... [325]
Formatted ... [326]
Formatted ... [327]
Formatted ... [328]
Formatted ... [329]

up a larger part of the genome, our defined RBP regulome is larger than previously thought and covered as much as 56.2 Mbp of the genome (Fig. 2A). In fact, the regulome is larger than the size of whole exome and only showed limited overlap with previous transcription-level annotations (Supplementary Fig. 2.). Furthermore, we found that the RBP regulome demonstrated noticeably larger conservation in two areas: higher cross-species conservation across all annotation categories (Fig. 2C) and significant enrichment in rare variants for most RBPs (Fig. 3A). These two sources of evidence support the notion that the RBP regulome is under strong purifying selection and carries out important biological functions. In addition, these results signify the necessity of computational tools to annotate and prioritize variants in the RBP regulome, which are under investigated.

By integrating a variety of regulator-, element-, and nucleotide-level features, we propose a entropy-based scoring frame, RADAR, to investigate impact of somatic and germline variants. The variant prioritization framework of RADAR contains two parts. First, by incorporating eCLIP, Bind-n-Seq, shRNA RNA-seq experiments with conservation and structural features, we built a pre-defined data context to quantify the baseline variant impact score. This approach is suitable for multiple-disease analysis or cases where no other prior information can be used. Applying this score to HGMD pathological variants highlighted many candidates that were solely discovered by RADAR and provided detailed explanation of the underlying disease-causing mechanism (Fig. 5). In addition to the baseline score, RADAR also allows user-specific inputs such as prior gene knowledge, patient expression, and mutation and survival profiles for a re-weighting process to highlight relevant variants in a disease-specific manner. As an example, we performed a breast cancer variant prioritization and score re-weighting scheme with user inputs in the well-known tumor suppressor gene TP53. Also results from somatic variants from several cancer types and showed that this scheme is able to identify relevant variants (Supplementary Fig. 7).

Formatted ... [337]
Deleted: large
Deleted: 2Mbp
Formatted ... [339]
Deleted: . It
Formatted ... [338]
Formatted ... [340]
Deleted: some
Formatted ... [341]
Deleted: .
Formatted ... [342]
Deleted: demonstrates
Formatted ... [343]
Deleted: from
Formatted ... [344]
Deleted: aspects. It does not only show
Formatted ... [345]
Deleted: under
Formatted ... [346]
Deleted: but also demonstrate
Formatted ... [347]
Deleted: .
Formatted ... [348]
Deleted: consistently showed
Formatted ... [349]
Deleted: carrying
Formatted ... [350]
Deleted: It signifies
Formatted ... [351]
Deleted: was previously
Formatted ... [352]
Formatted ... [353]
Formatted ... [354]
Deleted: ,
Formatted ... [355]
Deleted: ,
Formatted ... [356]
Deleted:
Formatted ... [357]
Deleted: proposed an
Formatted ... [358]
Deleted:
Formatted ... [359]
Deleted: impacts
Formatted ... [360]
Deleted: RADAR contains two parts in the
Formatted ... [361]
Deleted: structure
Formatted ... [362]
Deleted: It
Formatted ... [363]
Deleted:
Formatted ... [364]
Deleted: We applied
Formatted ... [365]
Formatted ... [366]
Formatted ... [367]
Formatted ... [368]
Formatted ... [369]
Formatted ... [370]
Formatted ... [371]
Formatted ... [372]
Formatted ... [373]

In summary, we have shown that RADAR is a useful tool for annotating and prioritizing post-transcriptional regulomes for RBPs, which has not been covered by most of the current variant functional impact interpretation tools. Our method also provides additional layers of information to current gene regulomes. Importantly, the RADAR scoring scheme can be used in conjunction with current transcriptional variant functional evaluation tools, such as Funseq, to evaluate variant impacts. Given the fast expanding collection of RBP binding profiles from additional cell types, we envision that RADAR can better tackle the functional consequence of mutations from both somatic and germline genomes.

Methods

eCLIP Data Processing and Quality Control

eCLIP is an enhanced version of the CLIP assay, and can be used to identify the binding sites of RBPs. We collected all available eCLIP experiments from the ENCODE data portal (encodeprojects.org), which included 178 experiments from K562 and 140 experiments from HepG2 cell lines, totaling 318 eCLIP experiments from all available ENCODE cell lines (released and processed by July 2017). These experiments targeted 112 unique RBP profiles. eCLIP data was processed per the ENCODE 3 uniform data processing pipeline. For each peak, we calculated the enrichment significance against a paired input, and then filtered the peaks with a significance flag of 1,000. We removed binding site locations containing blacklisted regions, including those on the genome with low sequencing depth or coverage. We further separated the peaks into coding regions and the noncoding regions (3'UTR, 5'UTR, 3'UTR extended, 5'UTR extended, and nearby intron regions).

Deleted: to annotate
Formatted ... [382]
Formatted ... [383]
Deleted: can serve as
Formatted ... [384]
Formatted ... [385]
Deleted: prioritize the
Formatted ... [386]
Deleted: It is
Formatted ... [387]
Deleted: able to provide
Deleted: More importantly
Formatted ... [388]
Formatted ... [389]
Deleted: some of the
Formatted ... [390]
Deleted: add independent information to jointly
Formatted ... [391]
Deleted: With
Formatted ... [392]
Deleted: of more RBPs
Formatted ... [393]
Deleted: more
Formatted ... [394]
Deleted: it
Formatted ... [395]
Deleted: more extensively
Formatted ... [396]
Deleted: 4
Formatted ... [397]
Formatted ... [398]
Formatted ... [399]
Deleted: 4.1
Formatted ... [400]
Formatted ... [401]
Formatted ... [402]
Formatted ... [403]
Formatted ... [404]
Deleted: crosslinking and immunoprecipitation (
Formatted ... [405]
Deleted:)
Formatted ... [406]
Deleted: is
Formatted ... [407]
Deleted: RNA binding proteins (
Formatted ... [408]
Deleted:).
Formatted ... [409]
Deleted:). There were
Formatted ... [410]
Deleted: The eCLIP peak calling method and proces ... [411]
Formatted ... [412]
Deleted: was calculated
Formatted ... [413]
Formatted ... [414]
Formatted ... [415]
Formatted ... [416]
Formatted ... [417]
Formatted ... [418]
Formatted ... [419]
... [421]
... [420]
Formatted ... [380]
Formatted ... [381]

Cross population conservations

To infer the selection pressure of a given region on the genome, we made use of germline variants from the 1,000 Genomes Project. The metric we used to represent negative selection pressure is the rare DAF. For a given region, i, containing rare variants n_i^r, and common variants n_i^c, the rare DAF is defined to be

$$r_{daf} = \frac{n_i^r}{n_i^r+n_i^c}$$

r_{daf} is usually confounded by its GC content. In order to correct for this bias, we first binned the genome into 500 base pair bins. Next, we estimated the average GC content within these 500 base pair bins and grouped bins with similar GC content to calculated the genome average. Finally, we took the rare DAF value and divided it by the GC content-corrected genome average to obtain a ratio. Regions with rare DAF ratios larger than 1 suggested an purifying selection pressure.

RBP Network Hubs

We termed the value associated with the number of RBPs that bind to a position the "hub size". To determine the actual relationship, we annotated the genome with its hub size on a base pair resolution. For both noncoding and coding regions, we estimated the selection pressure using the rare DAF ratio from germline variants within all regions showing an equal or larger hub size. We determined the rare DAF ratio by taking the rare DAF and dividing it by the corrected rare DAF, derived from evaluating the GC for regions with the same hub size and predicting the genomic rare DAF average. We found a cumulative relationship between rare DAFs and hotness, with a generally increasing trend. When the hotness exceeded 10, however, the lack of observations made it difficult to produce a reliable rare DAF. Therefore, we cut off the measure of rare DAFs at a maximum hub size of 10, corresponding to the top 1% of the data.

Formatted ... [424]
Formatted ... [426]
Deleted: to contextually annotating the genome wit ... [427]
Deleted: genetics data
Formatted ... [428]
Formatted ... [429]
Formatted ... [430]
Deleted: negative
Formatted ... [431]
Formatted ... [432]
Deleted: make
Formatted ... [433]
Deleted: 1000
Formatted ... [434]
Deleted: These germline variants consist of both co ... [435]
Formatted ... [436]
Deleted: derived allele frequency (rare
Formatted ... [437]
Deleted:).
Formatted ... [438]
Deleted: i
Formatted ... [439]
Deleted: r_i
Formatted ... [440]
Deleted:
Formatted ... [441]
Deleted: c_i,
Formatted ... [442]
Deleted: Rare DAF = r_i / (r_i + c_i) ... [443]
Formatted ... [444]
Formatted ... [445]
Deleted: bin
Formatted ... [446]
Deleted: estimate
Formatted ... [447]
Deleted: , which can range from 0% to 100%. We th ... [448]
Formatted ... [449]
Deleted: . Specifically, we establish 40 groups, usin ... [450]
Formatted ... [451]
Deleted: . For each of the 40 regions, i, we obtain a ... [452]
Formatted ... [453]
Deleted: 4.3.3 Negative selection pressure of RBP s ... [454]
Formatted ... [455]
Formatted ... [456]
Deleted: "hotness". Regions with more RBPs bindin ... [457]
Formatted ... [458]
Deleted: annotate
Formatted ... [459]
Deleted: hotness
Formatted ... [460]
Deleted: estimate
Formatted ... [461]
Comment [S 3]: Is this what you mean?
Deleted: to
Formatted ... [462]
Formatted ... [463]
Formatted ... [464]
Formatted ... [465]
Formatted ... [466]
Formatted ... [467]
Formatted ... [468]
Formatted ... [469]
Formatted ... [470]

Motif Analysis

We identified RBP motifs using DREME software (Version 4.12.0) directly from RBP peaks. In addition to the ENCODE eCLIP dataset, we utilized an *in vitro* RNA binding assay, RNA Bind-N-Seq (RBNS, Lambert et al. Molecular cell 54.5 (2014): 887-900), to characterize sequence and structural specificities of RBPs. We used RBNS motifs from 78 human RBPs to prioritize germline and somatic variants that could potentially disrupt an RNA-binding domain. Briefly, we called on RBNS motifs based on an enrichment Z-score cutoff of 3. Some RBPs had up to four motifs, which ranged from 5-mer to 9-mers.

To evaluate the functional importance of RNA-binding sites, we surveyed the mutational impact on RBP motifs. We called potential RBP motifs on high-confidence RBP peaks and evaluated the motif disruption power of each variant using a germline variant set (1,000 Genomes Project), a somatic variant set (30 types of cancer somatic single nucleotide variants), and HGMD (version 2015). We evaluated the motif breaking power using MotifTools (https://github.com/hoondy/MotifTools). We termed this the D-score, with D standing for disruptive-ness or deleterious-ness. We calculated the D-score based on the difference between sequence specificities in reference to an alternative sequence.

$$Dscore = motifscore_{ref} - motifscore_{alt} = -10\times \log_{10}\left(\frac{p_{ref}}{p_{alt}}\right)$$

We considered only positive D-scores, which denote a variant that decreases the likelihood that a transcription factor will bind the motif (motif-break), and ignored negative D-scores, which denote a variant that increases the likelihood that a transcription factor will bind the motif (motif-gain). For assessing D-score, we assumed a uniform nucleotide background, using a p-value threshold of 5e^{-2}. For each variant that affected multiple RBP binding profiles, we determined the max score.

Deleted: Many RBPs bind in similar locations acros ... [477]
Formatted ... [478]
Formatted ... [479]
Deleted: analysis
Formatted ... [480]
Deleted: 4.5.1 De novo discovery ... [481]
Formatted ... [482]
Formatted ... [483]
Formatted ... [484]
Deleted: 1000
Formatted ... [485]
Deleted: ,
Formatted ... [486]
Deleted: SNVs, Alexandrov et al., Nature 2013
Formatted ... [487]
Deleted: *** please confirm the version ***). Motif
Formatted ... [488]
Deleted: , which we labeled as D-score (D stands fo ... [489]
Formatted ... [490]
Deleted: was
Formatted ... [491]
Deleted: of
Formatted ... [492]
Formatted ... [493]
Formatted ... [494]
Deleted: only
Formatted ... [495]
Deleted: TF
Deleted: where
Formatted ... [496]
Formatted ... [497]
Deleted: TF to
Formatted ... [498]
Deleted: was assumed, and the
Formatted ... [499]
Deleted: was used
Formatted ... [500]
Deleted: were ***averaged***(we need to decide if
Formatted ... [501]
Deleted: average or max) over all D-scores.
Formatted ... [502]
Formatted ... [475]
Formatted ... [476]

INTUITION

Baseline Variant Scoring

Deleted: 4.6

Formatted: Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Formatted: Font color: Auto

Formatted: Font:14 pt, Font color: Auto

Deleted: 4.7 Regulatory Network Construction ... [503]

Comment [S 4]: I think it is confusing having this be "p" when you also use p values.

The three features that we considered in our baseline scoring scheme were conservation, network hubs, and motif disruption. In the first two scenarios, we calculated a value of *p*, corresponding to the number of germline variants falling in a given annotation divided by the total number of germline variants. For each of the three cases, the annotation was different. In the case of selection pressure, we naturally used an annotation equivalent to the binding sites from the eCLIP profiles. Each set of annotations was unique for each RBP, resulting in 112 different annotations and *p* values. In the scenario of network hubs, we calculated a *p* from the annotations corresponding to the number of RBPs binding at each base pair along the genome. For example, an annotation associated with a network hub of 1 would correspond to all the positions on the genome that only have 1 RBP binding. For each possible value for the network hub, we determined a value for *p* π. We used an entropy-based scoring scheme, with an input of *p* and an output between 0 and 1. We used the following scheme:

INTUITION [illegible]

$$Entropy\ score = 1 + p \log p + (1 - p) \log(1 - p)$$

In the case of the selection pressure score, we then reweighted the entropy score by the rare DAF by multiplying the rare DAF ratio, which takes into the GC bias associated with the rare DAF, and the entropy to give a final selection pressure score. When variants fell into RBP binding sites, the max selection pressure score was given to the variant's RADAR score.

For network hubs, we used a similar manner to determine a variant's network hub score. First, we intersected the variant with an annotation describing the number of RBPs that bind the position. This provided us with the value of how many RBPs bind. We then multiplied the entropy value associated with

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

this number (via the pre-calculated value of p) by the corresponding rare DAF ratio in order to reweight the final network hub score of a variant.

When a variant disrupted a motif, we used a slightly adjusted entropy scheme to score the variant. We first determined a function between entropy score and motif D-score using the following formula:

$$M_e^D = 1 + p^{\geq D} \log p^{\geq D} + (1 - p^{\geq D}) \log(1 - p^{\geq D})$$

Here M_e^D is the entropy score associated with a variant that breaks a motif with D-score, *D*. The $p^{\geq D}$ describes the number of germline variants breaking an RBP motif with a greater or equal D-score, *D*, divided by the total number of germline variants. Therefore, we can form a function between D-score and entropy score. However, in order to make the scheme more computationally efficient, we discretized the motif function between M_e^D and D-score by using D-scores at a step size of 0.5. To score a variant, we calculated a motif D-score, and used the nearest value from the discretized function for that associated motif to calculate the entropy score. Finally, we multiplied the entropy score by the rare DAF ratio of the RBP whose motif was broken.

[r TKC

RBP Regulatory Power from Linear Regression

RADAR allows inputs in addition to the pre-built context used to calculate the baseline variant score. We used DESeq to find tumor-to-normal differentially expressed genes (corrected p-val from DESeq <0.05).

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

For each RBP, we performed a linear regression to evaluate the individual regulatory potential on a set of its target genes.

Our x variable in the linear regression consisted of a vector of 1s and 0s, with vector length equal to the number of protein-coding genes. For each gene, the corresponding position in the vector x was equal to 0 if that gene was not in the regulatory network, and 1 if it was. The vector was rather sparse, containing many more 0s than 1s. The y variable consisted of a vector of differential expression of protein-coding genes. We determined these differential gene expression values for 17 different cancer types, allowing us to obtain 17 different regulatory potentials, depending on tissue type. We downloaded expression data from the TCGA data portal. We used count data from RNA-Seq in our analysis. The coefficient value from our linear regression determined the regulatory strength of an RBP, with the magnitude correlated to the strength of regulation and sign correlated to the direction of regulation (up or down regulate). To calculate the differential expression, we used DESeq2 (R Bioconductor package DESeq2 v3.5) due to its flexibility in allowing varying numbers of tumor and normal samples. We merged all cancer and normal samples into categories of cancer and tumor, respectively, to determine an appropriate differential expression. Therefore, each RBP network for each cancer type satisfied a linear regression, and we could infer the regulatory potential from the value of the coefficient. The associated p-value indicated the statistical significance that such a regulatory potential exists.

Formula

We performed a patient-wise regulatory potential linear regression, where we determined the differential expression as the individual expression fold change from a population mean. Each individual for a given cancer type was given a regulatory potential for each RBP, allowing for the regulatory potential of certain RBPs to serve as a prognosis marker. For each patient, we parsed the matching clinical XML data files for survival time. For patients who were alive, we used the number of days since the last follow up as a censored

Formatted ... [510]
Deleted: logistic
Formatted ... [507]
Formatted ... [508]
Deleted: explanatory
Deleted: , y,
Formatted ... [509]
Deleted: consists
Formatted ... [511]
Formatted ... [512]
Deleted:
Deleted: , xxx
Formatted ... [514]
Deleted: y is
Formatted ... [513]
Formatted ... [515]
Deleted: is
Formatted ... [516]
Deleted: is. This
Formatted ... [517]
Deleted: is
Formatted ... [518]
Deleted: x
Formatted ... [519]
Deleted: consists
Formatted ... [520]
Deleted: protein coding gene
Deleted: expressions. We determine
Formatted ... [521]
Formatted ... [522]
Deleted: 24
Formatted ... [523]
Deleted: 24
Formatted ... [524]
Deleted: Expression data is
Formatted ... [525]
Deleted: Data
Formatted ... [526]
Deleted: The
Formatted ... [527]
Deleted: is used in the analysis. The goal in differen ... [528]
Formatted ... [529]
Deleted: in the logistic
Formatted ... [530]
Deleted: in order
Formatted ... [531]
Deleted: indicate upregulation
Formatted ... [532]
Deleted: downregulation, respectively.
Formatted ... [533]
Deleted: is used,
Formatted ... [534]
Deleted: All
Formatted ... [535]
Formatted ... [536]
Formatted ... [537]
Formatted ... [538]
Formatted ... [539]
Formatted ... [540]
Formatted ... [541]
Formatted ... [542]
Formatted ... [543]

measure of survival time; however, we censored these patients to avoid an incorrect survival probability. We plotted survival curves with 95% confidence intervals.

Recurrence in Somatic Mutations

We prioritized variants that fell in elements (RBP binding sites) that were statistically enriched for somatic mutations. In order to do this, we first binned the genome using 1Mbp windows, and counted the number of somatic mutations in each window. This provided us a with a local mutation rate. Then, for each RBP binding site, we counted the number of somatic mutations, and compared it to the nearest local 1Mbp context using a one-sided binomial test. If a specific RBP binding site was enriched for somatic mutations, the variant falling in that site was given higher priority via the entropy scoring scheme.

Resource and Software Accessibility

We have made this RNA variant prioritization tool available as an open-source Python source at radar.gersteinlab.org. The website contains details on usage, examples, resources, and dependencies. We recommend a system with 10gb of RAM to avoid slowed performance for variant sets with a sample size below 1 million. We also provided a genome-wide pre-built RADAR score for every base pair on the genome (hg19 version of genome). Users can directly query the annotation and functional impact score from radar.gersteinlab.org. In addition, we released the RBP-gene regulatory network at radar.gersteinlab.org.

References

Deleted: . Survival curves are
Deleted: ,
Formatted ... [552]
Formatted ... [553]
Formatted ... [554]
Deleted: 4.9
Deleted: software accessibility
Formatted ... [555]
Formatted ... [556]
Formatted ... [557]
Formatted ... [558]
Deleted: ... [559]
Deleted: is made
Formatted ... [561]
Deleted:
Formatted ... [562]
Deleted: python
Formatted ... [560]
Formatted ... [563]
Deleted: xxx.
Formatted ... [564]
Deleted: A
Formatted ... [565]
Deleted: is recommended
Formatted ... [566]
Formatted ... [567]
Deleted: less than
Formatted ... [568]
Deleted:
Formatted ... [569]
Deleted: PASPort
Formatted ... [570]
Deleted: basepair
Formatted ... [571]
Deleted: xxxxxxxx (link). We also
Formatted ... [572]
Deleted: xxxxxxxx (link).
Formatted ... [573]
Formatted ... [574]
Formatted ... [575]
Formatted ... [576]
Formatted ... [550]
Formatted ... [551]

1 Croce, C. M. Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* **10**, 704-714, doi:10.1038/nrg2634 (2009).

2 Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L. & Almouzni, G. Epigenomics: Roadmap for regulation. *Nature* **518**, 314-316, doi:10.1038/518314a (2015).

3 Yang, G., Lu, X. & Yuan, L. LncRNA: a link between RNA and cancer. *Biochim Biophys Acta* **1839**, 1097-1109, doi:10.1016/j.bbagrm.2014.08.012 (2014).

4 Schmitt, A. M. & Chang, H. Y. Gene regulation: Long RNAs wire up cancer growth. *Nature* **500**, 536-537, doi:10.1038/nature12548 (2013).

5 Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829-845, doi:10.1038/nrg3813 (2014).

6 van Kouwenhove, M., Kedde, M. & Agami, R. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nat Rev Cancer* **11**, 644-656, doi:10.1038/nrc3107 (2011).

7 Swinburne, I. A., Meyer, C. A., Liu, X. S., Silver, P. A. & Brodsky, A. S. Genomic localization of RNA binding proteins reveals links between pre-mRNA processing and transcription. *Genome Res* **16**, 912-921, doi:10.1101/gr.5211806 (2006).

8 Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol* **3**, 195-205, doi:10.1038/nrm760 (2002).

9 Fu, X. D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**, 689-701, doi:10.1038/nrg3778 (2014).

10 Zheng, D. & Tian, B. RNA-binding proteins in regulation of alternative cleavage and polyadenylation. *Adv Exp Med Biol* **825**, 97-127, doi:10.1007/978-1-4939-1221-6_3 (2014).

Formatted: Indent: Left: 0", First line: 0"

Field Code Changed

Deleted: 2 Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827-841, doi:10.1093/nar/gks1284 (2013).
5

Deleted: 5

Deleted: 6

Deleted: 7

Deleted: 8

Deleted: 9

Deleted: 10

Deleted: 11

Deleted: 12

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

11 Fossat, N. *et al.* C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47. *EMBO Rep* **15**, 903-910, doi:10.15252/embr.201438450 (2014).

Deleted: 13

12 Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-1986, doi:10.1016/j.febslet.2008.03.004 (2008).

Deleted: 14

13 Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* **42**, D92-97, doi:10.1093/nar/gkt1248 (2014).

Deleted: 15

14 Blin, K. *et al.* DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* **43**, D160-167, doi:10.1093/nar/gku1180 (2015).

Deleted: 16

15 Anders, G. *et al.* doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* **40**, D180-186, doi:10.1093/nar/gkr1007 (2012).

Deleted: 17

16 Hu, B., Yang, Y. T., Huang, Y., Zhu, Y. & Lu, Z. J. POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res* **45**, D104-D114, doi:10.1093/nar/gkw888 (2017).

Deleted: 18

17 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).

Deleted: 19

18 Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell* **54**, 887-900, doi:10.1016/j.molcel.2014.04.016 (2014).

Deleted: 20

19 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913, doi:10.1101/gr.3577405 (2005).

20 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

Formatted: Indent: Left: 0", First line: 0"

Moved (insertion) [4]

21 Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480, doi:10.1186/s13059-014-0480-5 (2014).

Moved (insertion) [5]

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

22 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).

23 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

24 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

25 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

26 Garner, C. Confounded by sequencing depth in association studies of rare alleles. *Genet Epidemiol* **35**, 261-268, doi:10.1002/gepi.20574 (2011).

27 Xu, C., Nezami Ranjbar, M. R., Wu, Z., DiCarlo, J. & Wang, Y. Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller. *BMC Genomics* **18**, 5, doi:10.1186/s12864-016-3425-4 (2017).

28 Lu, Y. *et al.* Genetic variants cis-regulating Xrn2 expression contribute to the risk of spontaneous lung tumor. *Oncogene* **29**, 1041-1049, doi:10.1038/onc.2009.396 (2010).

29 Davidson, L., Kerr, A. & West, S. Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. *EMBO J* **31**, 2566-2578, doi:10.1038/emboj.2012.101 (2012).

30 Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15**, 469-479, doi:10.1038/nrg3681 (2014).

31 Pedersen, J. S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**, e33, doi:10.1371/journal.pcbi.0020033 (2006).

32 Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* **9**, e1002886, doi:10.1371/journal.pcbi.1002886 (2013).

Moved (insertion) [6]

Moved (insertion) [7]

Deleted: 21 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).
23 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).
38

Moved up [5]: Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480, doi:10.1186/s13059-014-0480-5 (2014).

Deleted: 35

Moved up [6]: Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).

Deleted: 36

Moved up [7]: Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

Deleted: 23 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).
38

Deleted: 37

Deleted: 38

Deleted: 39

Deleted: 40

Deleted: 41 Loh, T. J. *et al.* CD44 alternative splicing and hnRNP A1 expression are associated with the metastasis of breast cancer. *Oncol Rep* **34**, 1231-1238, doi:10.3892/or.2015.4110 (2015). ... [577]

Formatted: Indent: Left: 0", First line: 0"

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

33 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).

34 Consortium, G. T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

35 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

36 Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res* **43**, 8123-8134, doi:10.1093/nar/gkv803 (2015).

37 Lochovsky, L., Zhang, J. & Gerstein, M. MOAT: Efficient Detection of Highly Mutated Regions with the Mutations Overburdening Annotations Tool. *Bioinformatics*, doi:10.1093/bioinformatics/btx700 (2017).

38 Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat Med* **10**, 789-799, doi:10.1038/nm1087 (2004).

39 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).

40 Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* **24**, 52-60, doi:10.1016/j.gde.2013.11.014 (2014).

Acknowledgements

Author Contributions

Competing Financial Interests

Moved down [8]: S.

Moved down [9]: . *et al.*

Deleted: 45 Forbes,

Formatted: Font:Italic

Deleted: A

Deleted: The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10 11, doi:10.1002/0471142905.hg1011s57 (2008). ... [578]

Moved (insertion) [9]

Moved (insertion) [8]

Formatted: Font:Italic

Deleted: COSMIC: exploring the world's knowledge of somatic mutations in human

Deleted: .

Formatted: Indent: Left: 0", First line: 0"

Deleted: D805-811

Deleted: gku1075

Deleted: 47

Formatted: Indent: Left: 0", First line: 0"

Deleted: 48

Deleted: 49

Moved up [4]: Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

Deleted: 50

Deleted: 51 Singh, R. & Valcarcel, J. Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol* **12**, 645-653, doi:10.1038/nsmb961 (2005).

Comment [S 5]: Acknowledgements should be brief and should not include effusive comments. Grant or contribution numbers may be acknowledged.

Deleted:

Comment [S 6]: *Nature Communications* requires an author contributions statement as described in the Authorship section of our joint Editorial policies.

Comment [S 7]: A competing financial interests statement is required for all content of the journal. This statement will be published at the end of all papers, whether or not a competing financial interest is reported.

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

Figure Legends and Tables

Formatted: Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Formatted: Right: 0"

Page 2: [1] Deleted **Combine** **12/18/17 4:14:00 PM**

Dysregulation

Page 1: [2] Formatted **Combine** **12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [3] Formatted **Combine** **12/18/17 4:14:00 PM**

Right: 0"

Page 4: [4] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:26 pt

Page 4: [5] Formatted **Combine** **12/18/17 4:14:00 PM**

Heading 1, Left, None, Space Before: 0 pt, After: 0 pt

Page 4: [6] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:26 pt, Not Bold

Page 4: [7] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [8] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 4: [9] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [10] Deleted **Combine** **12/18/17 4:14:00 PM**

after (or simultaneously while) DNA has been transcribed to premature RNAs,

Page 4: [11] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [11] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [12] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [12] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [13] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [14] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [14] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [15] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [16] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [17] Deleted **Combine** **12/18/17 4:14:00 PM**

-6

Page 4: [17] Deleted **Combine** **12/18/17 4:14:00 PM**

-6

Page 4: [18] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [19] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [20] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [21] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [21] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [22] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 4: [23] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [24] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [25] Deleted **Combine** **12/18/17 4:14:00 PM**

-9

Page 4: [25] Deleted **Combine** **12/18/17 4:14:00 PM**

-9

Page 4: [26] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [27] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [28] Deleted **Combine** **12/18/17 4:14:00 PM**

10-14

Page 4: [28] Deleted **Combine** **12/18/17 4:14:00 PM**

10-14

Page 4: [29] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [30] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [30] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [31] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [31] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [31] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [31] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [32] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [32] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [33] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [34] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [35] Deleted	**Combine**	**12/18/17 4:14:00 PM**

19

Page 4: [35] Deleted	**Combine**	**12/18/17 4:14:00 PM**

19

Page 4: [36] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 4: [37] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 1: [38] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 5: [39] Deleted **Combine** **12/18/17 4:14:00 PM**

JZ2MG: whenever I talked about conservation, I am thinking about cross species conservation, not from the rare DAF stuff. Please pay attention to the red text below. I found it is confusing actually.

Page 5: [40] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [40] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [41] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [41] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [42] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [42] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [42] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [43] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [43] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [44] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [44] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [45] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 5: [45] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 5: [46] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 5: [46] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 5: [47] Formatted	Combine	12/18/17 4:14:00 PM

Heading 1, Left, None, Space Before: 0 pt, After: 0 pt

Page 5: [48] Formatted	Combine	12/18/17 4:14:00 PM

Font:26 pt

Page 5: [48] Formatted	Combine	12/18/17 4:14:00 PM

Font:26 pt

Page 1: [49] Formatted	Combine	12/18/17 4:14:00 PM

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [50] Formatted	Combine	12/18/17 4:14:00 PM

Right: 0"

Page 6: [51] Formatted	Combine	12/18/17 4:14:00 PM

Font:14 pt

Page 6: [52] Formatted	Combine	12/18/17 4:14:00 PM

Font:14 pt

Page 6: [53] Formatted	Combine	12/18/17 4:14:00 PM

Left

Page 6: [54] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [55] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [56] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [56] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [57] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [58] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [59] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [60] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [61] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [62] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [63] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [63] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [64] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [65] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 6: [66] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [67] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [67] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [68] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [69] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [70] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [70] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [71] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [72] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [73] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [74] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [75] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [76] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [77] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [77] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [77] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [77] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [78] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [79] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [80] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [80] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [81] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [82] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Page 6: [83] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:14 pt, Font color: Auto

Page 6: [84] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:14 pt, Font color: Auto

Page 6: [85] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 6: [86] Deleted **Combine** **12/18/17 4:14:00 PM**

Many researchers have pointed out

Page 6: [87] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 1: [88] Formatted **Combine** **12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [89] Formatted **Combine** **12/18/17 4:14:00 PM**

Right: 0"

Page 7: [90] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 7: [91] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [91] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [92] Deleted **Combine** **12/18/17 4:14:00 PM**

34-36

Page 7: [92] Deleted **Combine** **12/18/17 4:14:00 PM**

34-36

Page 7: [93] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [94] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [95] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [96] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [97] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [98] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [99] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [99] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [100] Deleted **Combine** **12/18/17 4:14:00 PM**

37,38

Page 7: [100] Deleted **Combine** **12/18/17 4:14:00 PM**

37,38

Page 7: [101] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [102] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [103] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [104] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [105] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [106] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [107] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [108] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [109] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [109] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [110] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [111] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [112] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [113] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [114] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [115] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [116] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [117] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 7: [118] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [119] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [119] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [120] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [121] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 7: [122] Deleted	**Combine**	**12/18/17 4:14:00 PM**

[40]. Hence, we used the enrichment of rare variants as a feature

Page 7: [123] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 1: [124] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [125] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Right: 0"

Page 8: [126] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [126] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [126] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [127] Moved from page 8 (Move #1) **Combine** **12/18/17 4:14:00 PM**

RNA secondary structures have been reported to affect every step of

Page 8: [128] Formatted **Combine** **12/18/17 4:14:00 PM**

Heading 3, Left, None, Space Before: 0 pt, After: 0 pt

Page 8: [129] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 8: [130] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 8: [131] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [132] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 8: [133] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [133] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [134] Deleted **Combine** **12/18/17 4:14:00 PM**

34,35,44

Page 8: [134] Deleted **Combine** **12/18/17 4:14:00 PM**

34,35,44

Page 8: [135] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [136] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [136] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [137] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [138] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [139] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [140] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [140] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [140] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [141] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [142] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [143] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [144] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [145] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [146] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [146] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 8: [146] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 8: [147] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 8: [148] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 8: [149] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 8: [150] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 8: [151] Deleted Combine 12/18/17 4:14:00 PM

2.2.3 Motif analysis for nucleotide impact

Mutations that change the RBP binding affinity may alter RBP regulation via loss-of-function effect. To quantify such impacts, we used the difference of position weight matrix (PWM) scores of the mutant allele against the reference allele.

Page 8: [152] Moved to page 9 (Move #2) Combine 12/18/17 4:14:00 PM

RADAR consists of two sources of motifs.

Page 8: [153] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 8: [154] Deleted Combine 12/18/17 4:14:00 PM

First, it has been reported that many of the RBPs' binding events *in vivo* can be captured by binding preferences *in vitro*. Hence, we used motifs reported by RNA Bind-n-Seq experiments from ENCODE. Then, we also used the same scheme on the *de novo* motifs discovered directly from the binding peaks using the default settings in DREME (details see methods). For each variant, the highest motif score from the above two sources is used to quantify the nucleotide effect.

2.2.4 Structure and context conservations

Page 8: [155] Formatted **Combine** **12/18/17 4:14:00 PM**

Heading 3, Left

Page 8: [156] Moved to page 8 (Move #1) **Combine** **12/18/17 4:14:00 PM**

RNA secondary structures have been reported to affect every step of

Page 8: [157] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 8: [158] Deleted **Combine** **12/18/17 4:14:00 PM**

the protein expression and RNA stability \cite{24821474}. We incorporated structure features predicted by Evofold, which uses phylogenetic stochastic context-free grammars to identify functional RNAs encoded in the human genome that are deeply conserved across species \cite{16628248}. We found after intersecting with conserved structure regions defined by Evofold, the RBP binding sites demonstrate significantly higher conversation, and hence we used the Evofold regions a feature in the baseline score.

Besides, cross-species sequence conservations have also been widely used as an important feature to discover regions with biological functions. For example, The Genomic Evolutionary Rate Profiling (GERP) score was developed to identify nucleotide level evolutional constraints by mapping human genome to other species. We used GERP score in our baseline RADAR framework to detect potentially deleterious mutations in the RBP regulome.

2.2.6 Highlight

Page 8: [159] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 8: [160] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 8: [161] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [161] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 8: [162] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 1: [163] Formatted **Combine** **12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [164] Formatted **Combine** **12/18/17 4:14:00 PM**

Right: 0"

Page 10: [165] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [166] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [166] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [167] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [168] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [169] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 10: [170] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [171] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [171] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [172] Deleted **Combine** **12/18/17 4:14:00 PM**

2.3.2 Prior knowledge of regulators, target genes, and variants

Variant Prioritization can be improved if

Page 10: [173] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 10: [174] Deleted **Combine** **12/18/17 4:14:00 PM**

of the regulators, target genes, and variants can be appropriately incorporated. For example,

Page 10: [175] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [176] Deleted **Combine** **12/18/17 4:14:00 PM**

splicing factor QKI has been reported by many reports to be associated with different cancer types and it is reasonable to further prioritize key variants that affect QKI regulation for cancer \cite{JZ2JL27841882 and others}. Besides, many databases have enumerated hundreds of cancer-associated genes

Page 10: [177] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [178] Deleted **Combine** **12/18/17 4:14:00 PM**

known to play critical roles in cancer. Cell proliferation and DNA repair-related genes are also important for cancer research. For other diseases, many GWAS studies have pointed out many risk genes. Furthermore, genes

Page 10: [179] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [180] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [181] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [182] Moved from page 11 (Move #3) **Combine** **12/18/17 4:14:00 PM**

In addition

Page 10: [183] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 10: [184] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [185] Formatted **Combine** **12/18/17 4:14:00 PM**

Left, None, Space Before: 0 pt, After: 0 pt

Page 10: [186] Deleted **Combine** **12/18/17 4:14:00 PM**

Hence, our RADAR framework allows users to input their disease-specific genes of interest to prune the candidate variant list further. Another example useful prior knowledge is

Page 10: [187] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [187] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [188] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [188] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [189] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 10: [190] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 10: [191] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 10: [192] Deleted	Combine	12/18/17 4:14:00 PM

events \cite{mutsigCV, larva, moat}. Taken the cancer variants as an

Page 10: [193] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 10: [194] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 1: [195] Formatted	Combine	12/18/17 4:14:00 PM

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [196] Formatted	Combine	12/18/17 4:14:00 PM

Right: 0"

Page 11: [197] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 11: [197] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 11: [198] Deleted	Combine	12/18/17 4:14:00 PM

, we can first define a local background mutation rate

Page 11: [199] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 11: [200] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 11: [201] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 11: [202] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:14 pt, Font color: Auto

Page 11: [203] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:14 pt, Font color: Auto

Page 11: [204] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [205] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Left

Page 11: [206] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [207] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [208] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [208] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [208] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [208] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [209] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [210] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [211] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [212] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [213] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [214] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [214] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [215] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [216] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [217] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [217] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [218] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [219] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [220] Formatted	**Combine**	**12/18/17 4:14:00 PM**

No Spacing, Left

Page 11: [221] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:12 pt, Not Bold, Font color: Auto

Page 11: [222] Formatted Table **Combine** **12/18/17 4:14:00 PM**

Formatted Table

Page 11: [223] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Not Bold, Font color: Auto

Page 11: [224] Formatted **Combine** **12/18/17 4:14:00 PM**

No Spacing, Left

Page 11: [225] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 11: [226] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 11: [227] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 11: [228] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 11: [229] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 11: [230] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 11: [231] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Not Bold, Font color: Auto

Page 11: [232] Formatted **Combine** **12/18/17 4:14:00 PM**

No Spacing, Left

Page 11: [233] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt, Font color: Auto

Page 11: [234] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:12 pt

Page 11: [235] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [236] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 11: [237] Formatted **Combine** **12/18/17 4:14:00 PM**

Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Page 11: [238] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:14 pt, Font color: Auto

Page 11: [239] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [240] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 11: [241] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [241] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [241] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [242] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [243] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [244] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [244] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 11: [245] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 1: [246] Formatted **Combine** **12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [247] Formatted **Combine** **12/18/17 4:14:00 PM**

Right: 0"

Page 12: [248] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [248] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [248] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [248] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [249] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [250] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [251] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [252] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [253] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [254] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [255] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [256] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [257] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [258] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [258] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [259] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [260] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [261] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [262] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [263] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [263] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [263] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [264] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [265] Formatted **Combine** **12/18/17 4:14:00 PM**

Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Page 12: [266] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:14 pt, Font color: Auto

Page 12: [267] Deleted **Combine** **12/18/17 4:14:00 PM**

2.6.1 Somatic variants associated with COSMIC genes and recurrence

Page 12: [268] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [269] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [270] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [271] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [272] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [272] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [273] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [274] Deleted **Combine** **12/18/17 4:14:00 PM**

47,48

Page 12: [274] Deleted	Combine	12/18/17 4:14:00 PM

47,48

Page 12: [275] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [276] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [277] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [278] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [279] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [279] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [280] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [281] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [281] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [282] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [283] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [284] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 12: [284] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [284] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [285] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [285] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [286] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [286] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [286] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 12: [287] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 1: [288] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [289] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Right: 0"

Page 13: [290] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Left

Page 13: [291] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [292] Deleted	**Combine**	**12/18/17 4:14:00 PM**

investigated (Fig S7).

In addition, because

Page 13: [293] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [294] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [294] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [295] Deleted Combine 12/18/17 4:14:00 PM

34-36

Page 13: [295] Deleted Combine 12/18/17 4:14:00 PM

34-36

Page 13: [296] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [297] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [298] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [299] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [299] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [299] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 13: [300] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [301] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [302] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [303] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [304] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [305] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [306] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [306] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [307] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [308] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [309] Formatted	Combine	12/18/17 4:14:00 PM

Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Page 13: [310] Formatted	Combine	12/18/17 4:14:00 PM

Font color: Auto

Page 13: [311] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [312] Formatted	Combine	12/18/17 4:14:00 PM

Left

Page 13: [313] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [314] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [315] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [316] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [317] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [318] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [319] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [320] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [321] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [322] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [322] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [322] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 13: [322] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [323] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [323] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [324] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [325] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [326] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [327] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [328] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [329] Formatted **Combine** **12/18/17 4:14:00 PM**

Heading 1, None, Space Before: 0 pt, After: 0 pt

Page 13: [330] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:26 pt

Page 13: [330] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:26 pt

Page 13: [331] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [332] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 13: [333] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 13: [334] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 1: [335] Formatted **Combine** **12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [336] Formatted **Combine** **12/18/17 4:14:00 PM**

Right: 0"

Page 14: [337] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [338] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [339] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [340] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [341] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [342] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [343] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [344] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [345] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [346] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [346] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [347] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [348] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [349] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [350] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [351] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [352] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [353] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [354] Formatted	Combine	12/18/17 4:14:00 PM

Left

Page 14: [355] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [356] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [357] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [358] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [359] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [359] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [359] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [360] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [361] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [361] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [362] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [363] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [364] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [365] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 14: [366] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [367] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [368] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [369] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [370] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [371] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [372] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [372] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [373] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [374] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [375] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [376] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [377] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 14: [378] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 14: [379] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 1: [380] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [381] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Right: 0"

Page 15: [382] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [383] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [384] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [385] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [386] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [386] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [386] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [386] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [387] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [388] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [389] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [390] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [391] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [392] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [392] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [393] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [394] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [395] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [396] Formatted Combine 12/18/17 4:14:00 PM

Font:11 pt

Page 15: [397] Formatted Combine 12/18/17 4:14:00 PM

Font:26 pt

Page 15: [398] Formatted Combine 12/18/17 4:14:00 PM

Heading 1, None, Space Before: 0 pt, After: 0 pt

Page 15: [399] Formatted	Combine	12/18/17 4:14:00 PM

Font:26 pt, Not Bold

Page 15: [400] Formatted	Combine	12/18/17 4:14:00 PM

Font color: Auto

Page 15: [401] Formatted	Combine	12/18/17 4:14:00 PM

Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Page 15: [402] Formatted	Combine	12/18/17 4:14:00 PM

Font:14 pt, Font color: Auto

Page 15: [403] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 15: [404] Formatted	Combine	12/18/17 4:14:00 PM

Left

Page 15: [405] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 15: [406] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 15: [407] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 15: [408] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 15: [409] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 15: [410] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 15: [410] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt

Page 15: [411] Deleted **Combine** **12/18/17 4:14:00 PM**

The eCLIP peak calling method and processing pipeline were developed by the laboratory of Gene Yeo at the University of California, San Diego (https://github.com/YeoLab/clipper, CLIP-seq cluster-identification algorithm on PMID: 24213538).

Page 15: [412] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [412] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [413] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [414] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [415] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [416] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [417] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [418] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [419] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 15: [420] Deleted **Combine** **12/18/17 4:14:00 PM**

4.2 Annotation

RNA binding proteins bind along the genome in a variety of contexts. Using eCLIP data, we can synthesize a genomic landscape of where RBPs bind. Raw peak signals from eCLIP data are translated into binding sites, using a peak caller specialized for eCLIP data. Generally, these RBPs having binding sites that correspond to about 150 bp, with many RBPs having well over 10,000 binding sites. Binding site locations containing blacklisted regions are removed. These include regions on the genome with low sequencing depth or coverage or […]. Despite filtering these blacklisted regions, over 99% of the binding locations are preserved. While the total number of base pairs corresponding to binding sites translates to a large number, compared to the scale of the genome it is still minute. Therefore, we annotate the genome, indicating at each position the set of RBPs that bind. This annotation set is known as the contextual annotations.

Page 15: [421] Commented	**SEN**	**12/16/17 7:30:00 PM**

Only one level of headings is included for Nature Communication

Page 1: [422] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [423] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Right: 0"

Page 16: [424] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Page 16: [425] Moved to page 10 (Move #3)	**Combine**	**12/18/17 4:14:00 PM**

In addition

Page 16: [426] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font color: Auto

Page 16: [426] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font color: Auto

Page 16: [427] Deleted	**Combine**	**12/18/17 4:14:00 PM**

to contextually annotating the genome with the preferential binding of RBPs, we also include a functional annotation – whether a specific position falls in the coding or noncoding region of the genome. The coding region consists of only the exons of protein coding genes. The noncoding region is further divided into 3’UTR, 5’UTR, 3’UTR extended, 5’UTR extended, and nearby intron regions. Coding and UTR annotations are retrieved from Gencode and UCSC, respectively. 3’UTR and 5’UTR extended regions consist of the 1000 base pairs downstream of the 3’UTR and 5’UTR regions, respectively. The nearby intron regions consist of the 100bp regions adjacent to each exon. While each of these region types are generally distinct, overlap is a possibility. Therefore, a hierarchy of which annotation takes precedence when annotation types overlap is established, from highest priority to lowest: coding, 3’UTR, 5’UTR, 3’UTR extended, 5’UTR extended, and nearby intron. Regions of the genome not classified by these annotations are labeled as “other” and may refer to other noncoding elements or blacklisted elements.

4.3 Inference of negative selection pressure from

Page 16: [428] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font color: Auto

Page 16: [429] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:14 pt, Font color: Accent 1

Page 16: [430] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Left

Page 16: [431] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [432] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [432] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [433] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [434] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [435] Deleted **Combine** **12/18/17 4:14:00 PM**

These germline variants consist of both common and rare variants. These variants are then classified into coding and noncoding variants. Coding variants fall in regions annotated as coding, while noncoding variants fall in regions annotated as noncoding Section (4.2). Noncoding variants are not further classified into noncoding element subgroups in order to maintain a large sample size for optimal statistical power in inferring negative selection pressure. The metric we use to

Page 16: [436] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [437] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [438] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [439] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [440] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [441] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [442] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [443] Deleted **Combine** **12/18/17 4:14:00 PM**

Rare DAF = r_i / (r_i + c_i)

Since we have further categorized both rare and common variants as coding and noncoding, we can obtain a coding and noncoding rare DAF for a given region as well. Finally, we take the rare DAF value and divide it by the GC content corrected genome average (Section 4.3.2) in order to obtain a ratio. Regions with rare DAF ratios larger than 1 suggest an above average negative selection pressure.

4.3.2 Rare DAF is confounded by GC content

Although negative selection pressure can be inferred from metrics such as rare DAF, it is not always accurate. In particular, the rare DAF of a region is severely

Page 16: [444] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 16: [445] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [446] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [447] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [448] Deleted **Combine** **12/18/17 4:14:00 PM**

, which can range from 0% to 100%. We then group

Page 16: [449] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [450] Deleted **Combine** **12/18/17 4:14:00 PM**

. Specifically, we establish 40 groups, using 2 percent intervals from 20 to 80 percent GC. Bins containing 0-20 and 80-100 percent GC content are ignored due to limited observations in these groups. For each of the 40 groups of 2% GC intervals, we associate a set of 500 base pair bins. Each of these sets are taken together to form a region, i, and the rare DAF is

Page 16: [451] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [452] Deleted **Combine** **12/18/17 4:14:00 PM**

. For each of the 40 regions, i, we obtain a rare DAF value, forming a discrete relationship between rare DAF and GC content. Using these discrete points, we fit a Gaussian kernel smoother with bandwidth of 10, resulting in a smoothed function between rare DAF and GC. This function serves as a way to estimate the genomic rare DAF given the GC content

Page 16: [453] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 16: [454] Deleted **Combine** **12/18/17 4:14:00 PM**

4.3.3 Negative selection pressure of RBP specific binding sites

We directly apply the method of determining a corrected rare DAF ratio to binding regions for a given RBP. The GC content of all binding sites for an RBP is estimated (from a genomic bigwig file), and using the derived smooth function between rare DAF and GC, a coding and noncoding rare DAF ratio is determined. For any given RBP a rare DAF ratio is used to measure the relative selection pressure of an RBP.

4.4 Co-binding and Hotness (need to brainstorm another title)

A natural extension to annotating locations based on the set of RBPs that preferentially bind, is to include the annotation of how many RBPs bind. The value associated with the number of RBPs that bind to a position is

Page 16: [455] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 16: [456] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 16: [457] Deleted **Combine** **12/18/17 4:14:00 PM**

“hotness”. Regions with more RBPs binding are deemed to be more “hot” than locations with fewer RBPs binding. We hypothesize that the hotness of a region and the selection pressure of the region demonstrate a positive relationship.

Page 16: [458] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [459] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [460] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [461] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [461] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [461] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [462] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [463] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [464] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [464] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [465] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [466] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [467] Formatted	Combine	12/18/17 4:14:00 PM

Font:11 pt, Highlight

Page 16: [468] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 16: [469] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 16: [469] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 16: [470] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 16: [471] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 16: [472] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 16: [473] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 16: [474] Deleted	**Combine**	**12/18/17 4:14:00 PM**

Furthermore, regions with hotness less than 5% of the data, equal to a hotness of less than 5, are deemed to not be hot, and are automatically given a 0 value in rare DAF ratio. The resulting discrete function is smoothed from hotness of 5 to 10. The function steps from 0 (from hotness of 1 to 4) to the rare DAF ratio at 5, and also maintains a constant rare DAF ratio for hotness values over 10 by rounding them down to 10.

Page 1: [475] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [476] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Right: 0"

Page 17: [477] Deleted	**Combine**	**12/18/17 4:14:00 PM**

Many RBPs bind in similar locations across the genome, and this is measured by their co-binding percent. The co-binding between two RBPs, A and B, is defined to be the maximum ratio between the peaks that intersect between A and B and the total number of peaks for A or B. Intersection is defined for greater than or equal to one base pair. Here, the maximum is taken in order to allow for a symmetric matrix in plotting a co-binding heatmap, resulting in only a unique possible result for clustering RBPs by similarity of co-binding. Using the co-binding ratio values between pairwise RBPs, a symmetric matrix is constructed and clustering is performed. The R function pvrect in package pvclust is used for clustering with an alpha value of 0.02 instead of 0.05 in order to avoid clusters with large numbers of RBPs (>6). The resulting clusters of RBPs with significance were found to follow patterns of functional co-binding found in literature.

4.5

Page 17: [478] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font color: Auto

Page 17: [479] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Page 17: [480] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:14 pt, Font color: Accent 1

Page 17: [481] Deleted	**Combine**	**12/18/17 4:14:00 PM**

4.5.1 De novo discovery

RBP motifs were found using DREME software (Version 4.12.0, http://meme-suite.org/tools/dreme, Timothy L. Bailey, "DREME: Motif discovery in transcription factor ChIP-seq data", Bioinformatics, 27(12):1653-1659, 2011.). De novo motif was called on a collection of significant eCLIP peaks.

4.5.2 Evaluating Motif Disruption with MotifTools

Page 17: [482] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [483] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 17: [484] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [484] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [485] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [486] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [487] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [488] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [489] Deleted **Combine** **12/18/17 4:14:00 PM**

, which we labeled as D-score (D stands for disruptive-ness or deleterious-ness), was evaluated

Page 17: [490] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [490] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [490] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [491] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [491] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [492] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [492] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [493] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [494] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 17: [495] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [495] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [496] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [497] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [498] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [498] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [499] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [500] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [501] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 17: [502] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 18: [503] Deleted **Combine** **12/18/17 4:14:00 PM**

4.7 Regulatory Network Construction

In order to construct a regulatory network of protein coding genes associated with a given RBP, we first identify which annotation is associated with which protein coding gene. The network we construct is undirected between protein coding genes and consists of a set of genes that a given RBP interacts with. To determine which genes the RBP interacts with, all binding sites of the RBP are intersected with all annotations (4.2). With the additional information of the associated gene given the annotation, we compile a list of all protein coding genes associated with the RBP. A unique list is determined and such a set of genes is determined to be the network of genes associated with that RBP. This is performed across each RBP in order to obtain a set of genes associated with each RBP.

4.8 RNA Binding Protein Prioritization

4.8.1 Logistic regression and regulation potential (add the DEseq analysis, have the software version clearly labeled)

To prioritize the RBPs we use a logistic regression approach. Our goal is to assess the regulatory potential (positive or negative) that the RBPs have on their respective gene associated targets. For each RBP we perform a logistic

Page 1: [504] Formatted **Combine** **12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [505] Formatted **Combine** **12/18/17 4:14:00 PM**

Right: 0"

Page 20: [506] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 20: [507] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [508] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 20: [509] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [510] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [511] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [512] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [512] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [513] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [514] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [515] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [516] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [517] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [518] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [519] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [520] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [521] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [522] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [523] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [524] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [525] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [525] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [525] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [526] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [527] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [528] Deleted	**Combine**	**12/18/17 4:14:00 PM**

is used in the analysis. The goal in differential expression is to allow for the detection of an extreme value for positive or negative

Page 20: [529] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [530] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [531] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [532] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [533] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [533] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [534] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [535] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [536] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [537] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [538] Formatted	**Combine**	**12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [539] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 20: [540] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [541] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [542] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [542] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [543] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [544] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [544] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [545] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [546] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [547] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [548] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 20: [549] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 1: [550] Formatted **Combine** **12/18/17 4:14:00 PM**

Position:Horizontal: Left, Relative to: Column, Vertical: In line, Relative to: Margin, Wrap Around

Page 1: [551] Formatted **Combine** **12/18/17 4:14:00 PM**

Right: 0"

Page 21: [552] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 21: [553] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt, Highlight

Page 21: [554] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [555] Formatted **Combine** **12/18/17 4:14:00 PM**

Font color: Auto

Page 21: [556] Formatted **Combine** **12/18/17 4:14:00 PM**

Heading 2, Left, None, Space Before: 0 pt, After: 0 pt

Page 21: [557] Formatted **Combine** **12/18/17 4:14:00 PM**

Font color: Accent 1

Page 21: [558] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:13 pt, Font color: Accent 1

Page 21: [559] Deleted **Combine** **12/18/17 4:14:00 PM**

This

Page 21: [560] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [561] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [562] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [563] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [564] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [565] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [566] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [567] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [568] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [569] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [570] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [571] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [572] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:11 pt

Page 21: [573] Formatted **Combine** **12/18/17 4:14:00 PM**

Left

Page 21: [574] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:26 pt

Page 21: [575] Formatted **Combine** **12/18/17 4:14:00 PM**

Heading 1, Left, None, Space Before: 0 pt, After: 0 pt

Page 21: [576] Formatted **Combine** **12/18/17 4:14:00 PM**

Font:26 pt, Not Bold

Page 24: [577] Deleted **Combine** **12/18/17 4:14:00 PM**

41 Loh, T. J. *et al.* CD44 alternative splicing and hnRNP A1 expression are associated with the metastasis of breast cancer. *Oncol Rep* **34**, 1231-1238, doi:10.3892/or.2015.4110 (2015).

42 Piton, A. *et al.* Analysis of the effects of rare variants on splicing identifies alterations in GABAA receptor genes in autism spectrum disorder individuals. *Eur J Hum Genet* **21**, 749-756, doi:10.1038/ejhg.2012.243 (2013).

43 Pala, M. *et al.* Population- and individual-specific regulatory variation in Sardinia. *Nat Genet* **49**, 700-707, doi:10.1038/ng.3840 (2017).

44

Page 25: [578] Deleted **Combine** **12/18/17 4:14:00 PM**

The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10 11, doi:10.1002/0471142905.hg1011s57 (2008).

46 Forbes, S. A.