TITLE: EXPLORING THE OVERLAP BETWEEN RHEUMATOID ARTHRITIS SUSCEPTIBILITY LOCI AND LONG NON-CODING RNA ANNOTATIONS

RUNNING TITLE: RHEUMATOID ARTHRITIS RISK LOCI AND LONG NON-CODING RNA

James Ding[1] (james.ding@manchester.ac.uk), Chenfu Shi[1], John Bowes[1,2], Stephen Eyre[1,2], Gisela Orozco[1,2]

[1]Centre for Genetics and Genomics Versus Arthritis. Division of Musculoskeletal and Dermatological Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, UK.

[2]NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, M13 9WL, UK.

ABSTRACT

Whilst susceptibility variants for many complex diseases, such as rheumatoid arthritis (RA), have been well characterised, the mechanism by which risk is mediated is still unclear for many loci. This is especially true for the majority of variants that do not affect protein-coding regions. lncRNA represent a group of molecules that have been shown to be enriched amongst variants associated with RA and other complex diseases, compared to random variants. In order to establish to what degree direct disruption of lncRNA may represent a potential mechanism for mediating RA susceptibility, we chose to further explore this overlap. By testing the ability of annotated features to improve a model of disease susceptibility, we were able to demonstrate a local enrichment of enhancers from immune-relevant cell types amongst RA susceptibility variants (log_2 enrichment 3.40). This was not possible for lncRNA annotations in general, however a small, but significant enrichment was observed for immune-enriched lncRNA (log_2 enrichment 0.867002). This enrichment was no longer apparent when the model was conditioned on immune-relevant enhancers (log_2 enrichment -0.372734), suggesting that direct disruption of lncRNA sequence, independent of enhancer disruption, does not represent a major mechanism by which susceptibility to complex diseases is mediated. Furthermore, we demonstrated that, in keeping with general lncRNA characteristics, immune-enriched lncRNA are expressed at low levels that may not be amenable to functional characterisation.

INTRODUCTION

In keeping with other complex diseases, rheumatoid arthritis (RA) susceptibility loci are mainly non-coding, with relatively few variants having a potential impact upon the coding sequence for a protein [1,2]. Enhancers have been identified as likely to mediate disease susceptibility at many loci. Evidence to support this generalization includes demonstrated effects of genome wide association study (GWAS) variants on enhancers at individual loci [3], as well as an enrichment of RA GWAS variants amongst enhancers from relevant cell types [4]. Alternative non-coding elements, such as long non-coding RNA (lncRNA) may also play a role in mediating the increased risk associated with non-coding variants.

lncRNA are a heterogeneous class of molecules that are defined based on a lack of protein-coding potential and a minimum transcribed length of 200 nucleotides. Whilst discrete subcategories exist, such as long intergenic non-coding RNAs, promoter associated-lncRNAs or antisense lncRNA with some discriminatory characteristics, including genomic context, overlapping chromatin marks, length and structure [5], it can still be difficult to discriminate a genuine lncRNA annotation from a product of spurious transcription. Many individual lncRNA have been functionally characterised, with gene regulation featuring frequently amongst the wide variety of roles described. One important subcategory was identified following observations of RNA polymerase II recruitment and transcription at enhancers [6]. Often described as enhancer derived RNA (eRNA), expression of these transcripts is highly correlated with enhancer activity

[7] and increasing evidence suggests that these ncRNA may contribute towards enhancer function, although the precise mechanism is still unclear [8].

There is some evidence to suggest that GWAS susceptibility variants are enriched amongst lncRNA [9,10]. However, using conventional methods of determining whether annotations overlap more than can be expected by chance it is difficult to appropriately account for confounding factors, such as chromosomal compartments or chromatin accessibility. Using the *ab initio* MiTranscriptome assembly (58,648 lncRNA), which was generated using a large collection of RNA sequencing libraries, GWAS SNPs were demonstrated to be enriched in lncRNA, compared to other SNPs tested for in GWAS analyses [9]. This enrichment was also observed using either GWAS SNPs or probabilistically identified causal SNPs (PICS) and lncRNA from the functional annotation of the human genome (FANTOM) cap-analysis gene expression (CAGE) associated transcriptome (CAT) assembly (27,919 lncRNA), generated using CAGE datasets in combination with existing assemblies [10]. Using the tissue specific nature of the FANTOM CAT annotation it was also possible to demonstrate that this enrichment was markedly higher when testing specifically for an enrichment of immune-relevant GWAS PICS in immune-expressed lncRNA transcripts.

Despite these studies, the relevance of lncRNA to the study of individual complex diseases, such as RA, remains unclear. This is especially true given the overlap in genomic locality and function between enhancers and lncRNA. We chose to investigate the overlap between lncRNA annotations, enhancer annotations and GWAS SNPs associated with RA susceptibility, with the aim of establishing to what degree the

direct disruption of lncRNA by RA-associated variants may contribute to the mediation of disease risk. Central to our investigation is the use of the fgwas algorithm [11], which tests the ability of individual annotations to improve a probabilistic model of disease susceptibility, constructed using GWAS summary statistics. Using this method a local enrichment is estimated, that takes into consideration the non-random distribution of potentially confounding genomic features. In addition, it is possible to model multiple traits and establish the degree to which they are independently predictive.

MATERIALS AND METHODS

ENRICHMENT TESTING

Enrichment testing was performed using RA summary statistics [1] and fgwas v0.3.6 [11], with the "-cond" option called for conditional analyses. Chromatin state data was obtained from the Roadmap Epigenomics project [12], with the expanded, 18-state model used for all 98 corresponding epigenomes. The 18-states were combined to form four exclusive annotations as follows: TSSs: Active TSS (1), Flanking TSS (2), Flanking TSS upstream (3), Flanking TSS downstream (4), Bivalent/poised TSS (14). Transcription: Strong transcription (5), Weak Transcription (6). Enhancers: Genic enhancer 1 (7), Genic enhancer 2 (8), Active enhancer 1 (9), Active enhancer 2 (10), Weak enhancer (11), Bivalent enhancer (15). Repressed chromatin: ZNF genes and repeats (12), Heterochromatin (13), Repressed polycomb (16), Weak repressed polycomb (17), Quiescent/low (18). The following lncRNA datasets were interrogated: Lncipedia v5.2 [13], miTranscriptome v2 [9] and FANTOM CAT (robust) [10]. Immune-

relevant enhancers were defined as genomic regions annotated as enhancers in cell-types defined as originating from "Blood and T cell" or "HSC and B cell" by the Roadmap Epigenomics project. The definition of Immune-enriched lncRNA is based on the underlying sample ontology and was wholly adopted from Hon et al. requiring: detection in at least 50% of immune-relevant samples, 5 x higher expression in immune-relevant samples than in other samples and $P < 0.05$ in a one-tailed Mann–Whitney rank sum test [10].

EXPRESSION PROFILING

Raw RNA-seq reads were downloaded from the Roadmap Epigenomics project for primary T-helper cells (SRA accession SRR644513 and SRR643766) [12]. Reads were then filtered for quality, adapter content and polyA tails using fastp version 19.7, with default settings and polyX tail trimming enabled. Transcripts were quantified using Salmon version 13.1 [14] using suggested settings ("quant" mode and "–validatemappings") and using the reference index generated from the FANTOM CAT robust database [10]. Transcripts quantifications (reported as Transcripts per million, TPM) were then remapped to genes and summed for each gene. CAGE transcript counts per million (CPM) were taken from FANTOM CAT for T-helper cells (CL_0000084_T_cell), as published [10].

Statistical difference between the distributions of immune-enriched lncRNA and mRNA abundance was established using a two-sided Welch's t-test, with no assumption of equal variance. This test assumes a normal distribution of the mean. Using the Mann-

120 Whitney u-test, which does not require this assumption, it was not possible to
121 estimate a p-value as it is too close to zero.

122 ## Results

123 ### Enhancer annotations from immune-relevant cell types are enriched amongst RA
124 susceptibility variants

125 An enrichment of RA PICS has previously been demonstrated amongst *cis*-regulatory
126 elements that are active in T-helper cells and lymphoblastoid cells [4], using data from
127 the high density Immunochip custom SNP array. In order to establish confidence in the
128 ability of fgwas to identify similar enrichments we sought to validate this evidence of
129 enrichment using a more inclusive approach that incorporates the probability of
130 association for all SNPs tested in the most recent RA GWAS meta-analysis [1]. In order
131 to achieve this, chromatin state data taken from the Roadmap Epigenomics project
132 [12] was incorporated in a model of RA susceptibility. Using enrichment estimates
133 generated using the expanded 18-state model, it is possible to discern an enrichment
134 of certain chromatin states, such as genic enhancers, active enhancers and weak
135 enhancers in immune-relevant cell types, such as B cells, T cells and monocytes (panel
136 A in S1 Fig), however the confidence intervals (CIs) for estimates are broad for many
137 states, likely due to a reduced abundance of these annotations (panel B and C in S1
138 Fig). The size of CIs was improved by combining states to generate four more easily
139 interpretable annotations (enhancers, transcription start sites (TSSs), transcription,
140 and repressed chromatin; Fig 1). In keeping with our understanding of RA, the highest
141 level of enrichment was observed for enhancer annotations in immune-relevant cell

types, with regulatory T cells showing the highest enrichment (log_2 enrichment 3.17, 95% CI 2.58; 3.75015). In immune-relevant cell types TSSs were also enriched, whilst repressed chromatin was depleted (Fig 1).

LNCRNA ANNOTATIONS SHOW NEGLIGIBLE ENRICHMENT AMONGST RA SUSCEPTIBILITY VARIANTS

We applied fgwas to test for an enrichment of lncRNA amongst RA susceptibility variants, using lncipedia, a large lncRNA database curated from a number of sources [13], the MiTranscriptome assembly [9], and the FANTOM CAT assembly [10]. Using fgwas, MiTranscriptome lncRNA genes show a level of depletion amongst RA susceptibility variants comparable to repressed chromatin. MiTranscriptome lncRNA exons and both genes and exons from either FANTOM CAT or lncipedia all show negligible enrichment (Fig 2A).

LNCRNA ANNOTATIONS WITH ENRICHED EXPRESSION IN IMMUNE-RELEVANT CELLS ARE SUBTLY ENRICHED AMONGST RA SUSCEPTIBILITY VARIANTS

Uniquely, FANTOM CAT transcripts are associated with tissue specific expression data. As in the original FANTOM CAT publication, we took advantage of this additional information, to test for an enrichment of lncRNA whose expression is enriched in immune-relevant cell types amongst RA susceptibility loci. This approach demonstrated a subtle enrichment of lncRNA genes (log_2 enrichment 0.867, 95% CI 0.0554; 1.57) and similar level of enrichment for their exons, albeit with an increased confidence interval (log_2 enrichment 0.799, 95% CI 2.30; 1.94). FANTOM CAT mRNA annotations, whose expression is enriched in immune-relevant cell types, were included in order to provide a comparison. Genic mRNA annotations showed a similar

level of enrichment as lncRNA genes, with mRNA exons exhibiting slightly higher enrichment (Fig 2B).

THE SUBTLE ENRICHMENT OF IMMUNE-ENRICHED LNCRNA OBSERVED IS NOT INDEPENDENT OF IMMUNE-RELEVANT ENHANCER ANNOTATIONS

Given the strong enrichment of immune-relevant enhancers amongst RA susceptibility loci and the established overlap between lncRNA annotations and enhancers, we were interested to investigate the independence of these variables using fgwas. In this conditional analysis, a residual enrichment of immune-enriched FANTOM CAT annotations was tested after the enrichment of immune-relevant enhancers (log_2 enrichment 3.40, 95% CI 2.54; 4.58) was accounted for (Fig 3Fig 3). Interestingly, both lncRNA and mRNA annotations no longer show significant enrichment, indicating that once enrichment of susceptibility variants in enhancers has been accounted for no remaining enrichment of mRNA or lncRNA is apparent.

IMMUNE-ENRICHED LNCRNA ARE EXPRESSED AT LOW LEVELS IN RA RELEVANT CELL TYPES

lncRNA are generally considered to exhibit low expression levels and high tissue specificity that can make them difficult to study using conventional methods. Given their enrichment for RA susceptibility variants we were interested to establish whether this description applied to FANTOM CAT immune-enriched lncRNA in an RA relevant cell type. Using randomly primed Roadmap Epigenomics RNA-seq data from primary T-helper cells, the distribution of expression levels for FANTOM CAT immune-enriched lncRNA is significantly lower than that of FANTOM CAT immune-enriched mRNA ($p = 4.46 \times 10^{-34}$, Fig 4A median lncRNA transcripts per million reads (TPM); 0.257, vs 124

for mRNA). The same is true using less-conventional expression profiling methods, such as those employed by FANTOM CAT, which offer improved sensitivity for the detection of transcripts of low abundance ($p = 1.61 \times 10^{-24}$, Fig 4B median lncRNA counts per million reads (CPM); 1.43, vs 71.7 for mRNA). 90% of immune-enriched lncRNA have abundance lower than 85.2% of immune-enriched mRNA in Roadmap Epigenomics RNA-seq data or 75.6% of immune-enriched mRNA in FANTOM CAT CAGE data.

Discussion

By incorporating both cell-type specific enhancer and lncRNA annotations into a probabilistic model of RA susceptibility it is possible to demonstrate their respective levels of enrichment amongst RA susceptibility variants. Whilst previous studies have demonstrated an enrichment of lncRNA compared with randomly shuffled annotations these analyses fail to take into consideration the complex organisation of the genome and are easily confounded by alternative features. In our analysis, which incorporated lncRNA from various databases, it was only possible to demonstrate a subtle enrichment of lncRNA whose expression was previously identified as being enriched in relevant cell-types.

By conditioning a model of RA susceptibility on enhancer annotations from immune-relevant cell types it is possible to test the independence of additional features. This demonstrated that the subtle enrichment observed for immune-relevant lncRNA is entirely explained by immune-relevant enhancer annotations. Interestingly, the same is true of mRNA annotations, indicating that in both instances the primary influence of

susceptibility variants is in affecting non-coding regulatory elements, with any effect on mRNA and lncRNA being secondary and/or indirect. This suggests that the majority of genetic variance in disease susceptibility is mediated through disruption of regulatory elements and not through direct disruption of transcript sequences. This observation is in keeping with those made previously, relating to the minimal overlap between the coding sequence of mRNA and RA susceptibility variants. It is also consistent with the existence of well-characterised effects of RA susceptibility variants on coding regions, such as for the HLA proteins [15], which convey a significant, but not exhaustive proportion of risk.

Similarly, this analysis does not rule out the relevance of lncRNA at individual loci, in fact it is worth noting that there is evidence to suggest that C5T1lncRNA may mediate risk at a RA risk locus located at chromosomal position 9q33.2, with RA associated variants falling within a C5T1lncRNA exonic region [16]. Our analysis, however, precludes an effect of RA susceptibility variants on the transcribed sequence of lncRNA independent of enhancer disruption. Sequence-specific functions of lncRNA are, therefore, unlikely to mediate a significant proportion of the risk-modifying effect associated with RA susceptibility variants. Furthermore, the distribution of FANTOM-CAT immune-enriched lncRNA expression levels in primary T-helper cells highlight the difficulties associated with studying lncRNA, whose expression is typically very low and highly cell-type specific.

The analyses performed are specific to RA, however it is assumed that similar results would be reached using GWAS data, enhancer annotations and lncRNA annotations

230 relevant to other diseases. This would, therefore, suggest that the enrichment of
231 lncRNA annotations amongst GWAS variants observed by others [9,10], may result
232 from a high degree of overlap between regulatory features such as enhancers and
233 lncRNA and other confounding features, such as active and inactive genomic
234 compartments. As a generalisation, when it comes to functional characterization of
235 variants associated with complex genetic disorders, sequence-specific lncRNA
236 functions is unlikely to represent an attractive area for study; lncRNA are not
237 independently enriched amongst such variants and are difficult to study, due to low
238 expression levels. Despite these results, dysregulation of lncRNA expression could still
239 play a role in RA and similar diseases, with disease associated variants affecting
240 regulatory elements, such as enhancers that control lncRNA expression.

241 fgwas represents a useful tool for studying the enrichment of different features
242 amongst susceptibility variants, especially when used in combination with Roadmap
243 Epigenomics data in order to identify cell-types and tissues that are relevant for
244 disease susceptibility. Whilst using this tool we observed that the number and size of
245 annotations that are tested have a strong influence, both on the confidence with
246 which any enrichment is estimated, as well as on the extent of that enrichment. It is
247 likely that this may explain some of the subtle differences observed between different
248 lncRNA databases.

249 Our analyses highlight the caveats associated with inferring functional relevance for a
250 given feature, based purely on the observation of enrichment over a genomic
251 background, as well as the care that must be taken when attempting to interpret such

252 enrichments. By deriving enhancer and lncRNA annotations from entirely different
253 sources we have tried to ensure that the demonstrated dependence is not self-
254 fulfilling, as it may have been if we defined immune-relevant lncRNA based on
255 underlying chromatin states.

256 In conclusion, using fgwas and Roadmap Epigenomics chromatin state data it is
257 possible to identify cell types and chromatin states of relevance to complex diseases,
258 such as RA. In the case of RA this is predominantly enhancers and transcription start
259 sites from immune-relevant cell types. It is also possible to test the association of
260 alternative features and establish their independence from chromatin states. Here, a
261 previously described enrichment of lncRNA amongst GWAS susceptibility loci was
262 explored for RA. Immune-enriched lncRNA from the FANTOM-CAT database were
263 found to be enriched amongst RA susceptibility loci, however, this enrichment was not
264 apparent when chromatin-state data was taken into account.

265 Our results suggest that regulatory elements, such as enhancers, are likely to mediate
266 the vast majority of variance in risk associated with RA and other complexes diseases,
267 with no substantial independent contribution being made by direct disruption of
268 lncRNA sequences. Because of this, and the difficulties associated with detecting
269 transcripts of such low abundance, sequence-specific lncRNA function does not
270 represent the most attractive area for study with respect to RA susceptibility, except in
271 the case of in depth characterisation of individual loci.

272 **ACKNOWLEDGEMENTS**

The authors would like to acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester.

References

1. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2013;506: 376–381. doi:10.1038/nature12873

2. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet. 2012;44: 1336–40. doi:10.1038/ng.2462

3. Simeonov DR, Gowen BG, Boontanrart M, Roth TL, Gagnon JD, Mumbach MR, et al. Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature. 2017;549: 111–115. doi:10.1038/nature23875

4. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2014;518: 337–43. doi:10.1038/nature13835

5. Jarroux J, Morillon A, Pinskaya M. History, discovery, and classification of lncRNAs. Advances in Experimental Medicine and Biology. Springer, Singapore; 2017. pp. 1–46. doi:10.1007/978-981-10-5203-3_1

6. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465: 182–7. doi:10.1038/nature09033

294 7. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al.
295 An atlas of active enhancers across human cell types and tissues. Nature.
296 2014;507: 455–61. doi:10.1038/nature12787

297 8. Lam MTY, Li W, Rosenfeld MG, Glass CK. Enhancer RNAs and regulated
298 transcriptional programs. Trends Biochem Sci. 2014;39: 170–82.
299 doi:10.1016/j.tibs.2014.02.007

300 9. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of
301 long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47: 199–
302 208. doi:10.1038/ng.3192

303 10. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, et al. An
304 atlas of human long non-coding RNAs with accurate 5 ends. Nature. 2017;543:
305 199–204. doi:10.1038/nature21374

306 11. Pickrell JK. Joint analysis of functional genomic data and genome-wide
307 association studies of 18 human traits. Am J Hum Genet. 2014;94: 559–73.
308 doi:10.1016/j.ajhg.2014.03.004

309 12. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al.
310 Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:
311 317–330. doi:10.1038/nature14248

312 13. Volders P-J, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, et al.
313 LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic
314 Acids Res. 2019;47: D135–D139. doi:10.1093/nar/gky1031

14. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419. doi:10.1038/nmeth.4197

15. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee H-S, Jia X, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet. 2012;44: 291–6. doi:10.1038/ng.1076

16. Messemaker TC, Frank-Bertoncelj M, Marques RB, Adriaans A, Bakker AM, Daha N, et al. A novel long non-coding RNA in the rheumatoid arthritis risk locus TRAF1-C5 influences C5 mRNA levels. Genes Immun. 2015; doi:10.1038/gene.2015.54

FIGURE CAPTIONS

FIG 1. ENRICHMENT OF CHROMATIN STATE GROUPS AMONGST RA SUSCEPTIBILITY VARIANTS FOR 98 CELL TYPES

Estimates for enrichment of combined chromatin state groupings are illustrated for all 98 cell types annotated within the Roadmap Epigenomics 18-state model. Cell-types are ordered and coloured according to the clustering established by the Roadmap Epigenomics project, with immune-relevant cell types coloured green. Estimates and confidence intervals are clipped at axis limits, where applicable.

FIG 2. ENRICHMENT OF LNCRNA ANNOTATIONS AMONGST RA SUSCEPTIBILITY VARIANTS

Estimates for the enrichment of genic (black circle) and exonic (grey diamond) annotations from a variety of lncRNA containing databases, including 95% confidence intervals (A). Separate estimates are included for annotations identified as exhibiting enriched expression in immune-relevant cells (B).

FIG 3. ENRICHMENT OF IMMUNE-ENRICHED LNCRNA AMONGST RA SUSCEPTIBILITY VARIANTS AFTER CONDITIONING ON CHROMATIN STATE DATA

The influence of immune-relevant enhancers (red circle) was fixed in a probabilistic model of RA susceptibility to determine whether the subtle enrichment of FANTOM CAT immune-enriched lncRNA or mRNA adds any additional predictive information and is therefore independently enriched. Genic (black circle) and exonic (grey diamond) annotations were both tested.

As may be expected given the magnitude of enrichments observed, after accounting for the effect of immune-enriched FANTOM CAT annotations, the residual enrichment of immune-relevant enhancers is not dramatically reduced (S2 Fig).

FIG 4. DISTRIBUTION OF IMMUNE-ENRICHED LNCRNA AND MRNA EXPRESSION LEVELS IN PRIMARY T-HELPER CELLS

Staggered bars are used to illustrate the proportion of transcripts whose expression falls in bins of 25 million transcripts, or counts, in Roadmap Epigenomics RNA-seq data (A) and FANTOM CAT CAGE data (B), respectively.

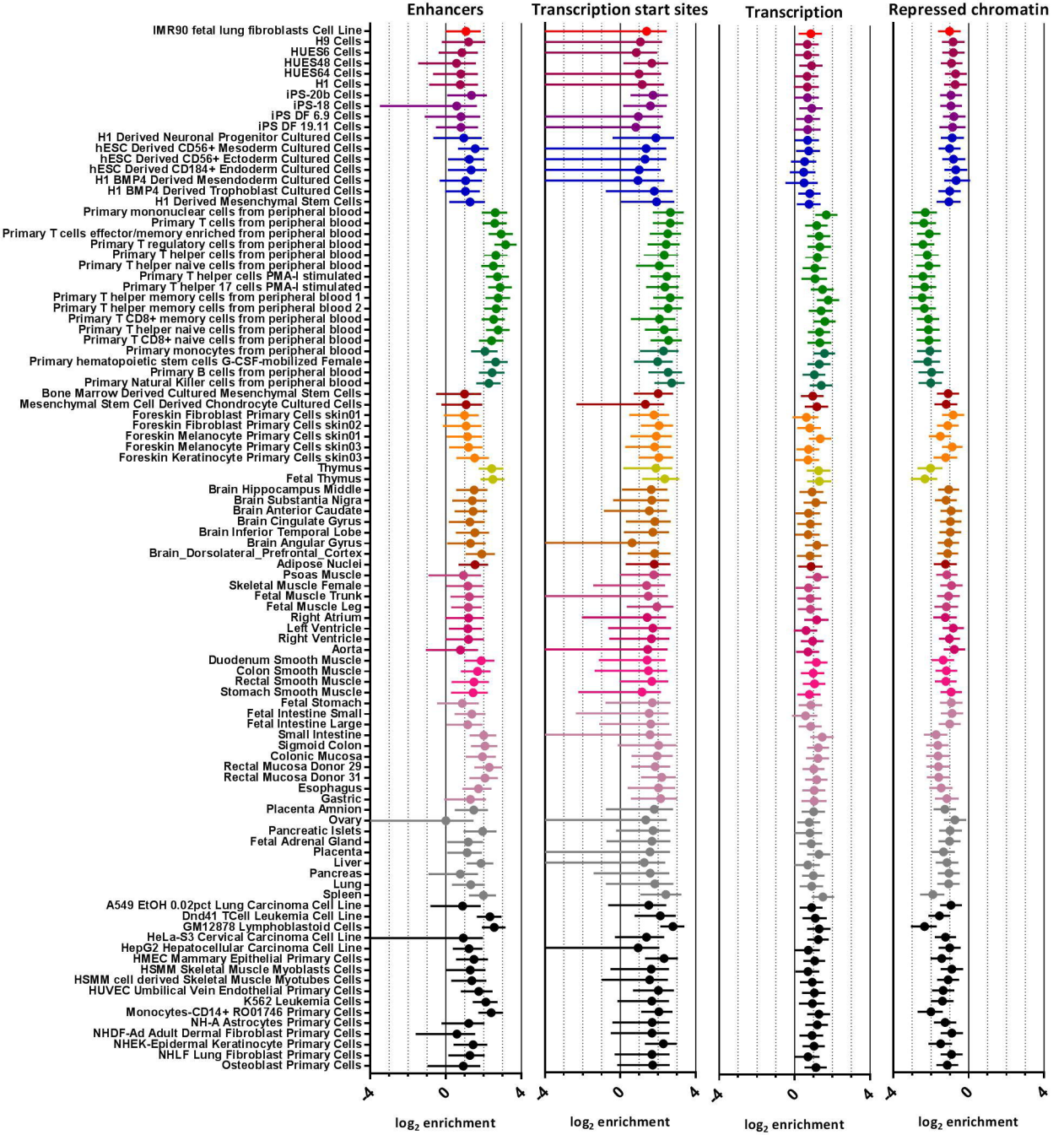
SUPPORTING INFORMATION CAPTIONS

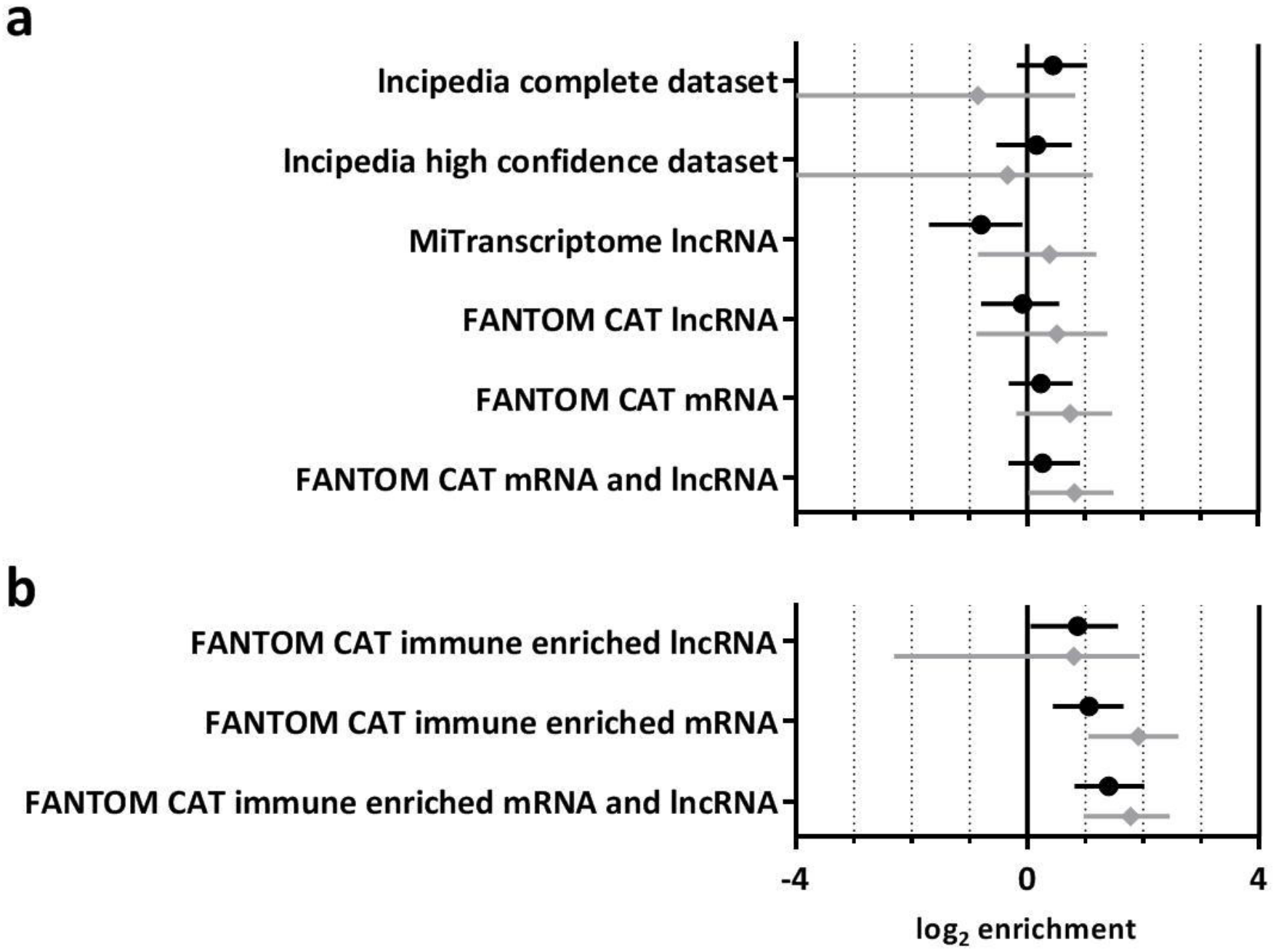
S1 FIG. OF CHROMATIN STATE ANNOTATIONS AMONGST RA SUSCEPTIBILITY VARIANTS

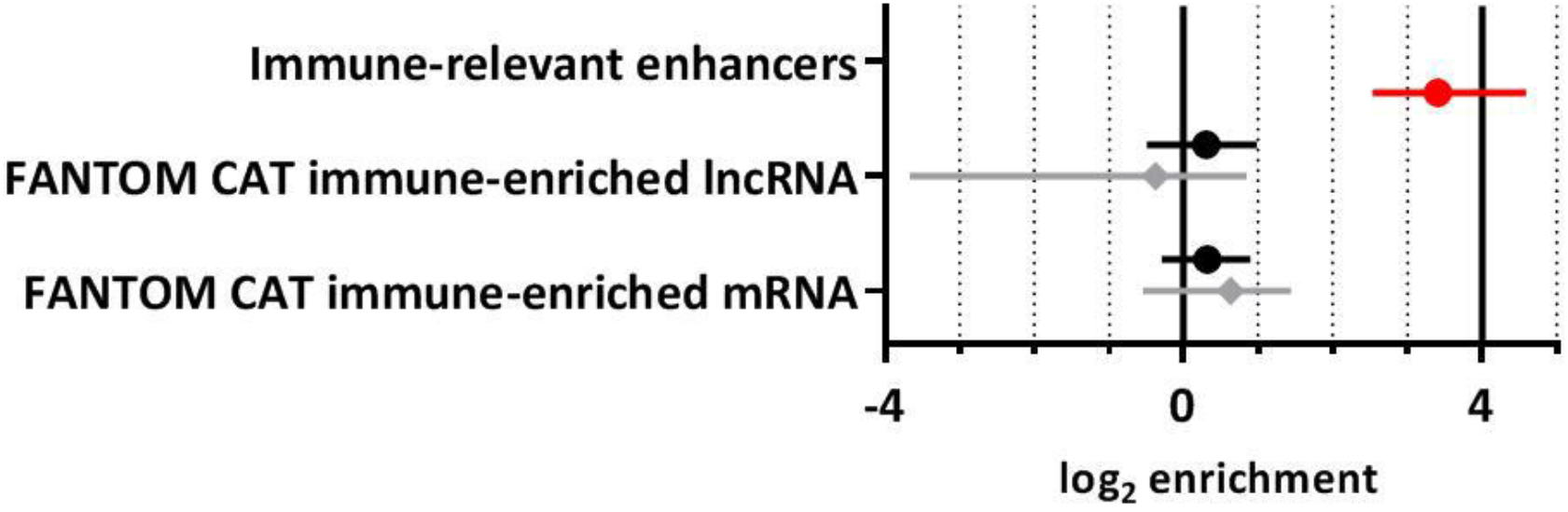
Estimates for enrichment of individual states are illustrated for 98 cell types using the Roadmap Epigenomics 18-state model (a). Similar states were grouped into four groups for all immune-relevant primary cell-types (b), as individual states often gave very broad 95% confidence intervals (c). Cell-types are ordered according to the clustering established by the Roadmap Epigenomics project, with chromatin states reordered according to their subsequent grouping. Estimates and confidence intervals are clipped at axis limits, where applicable.

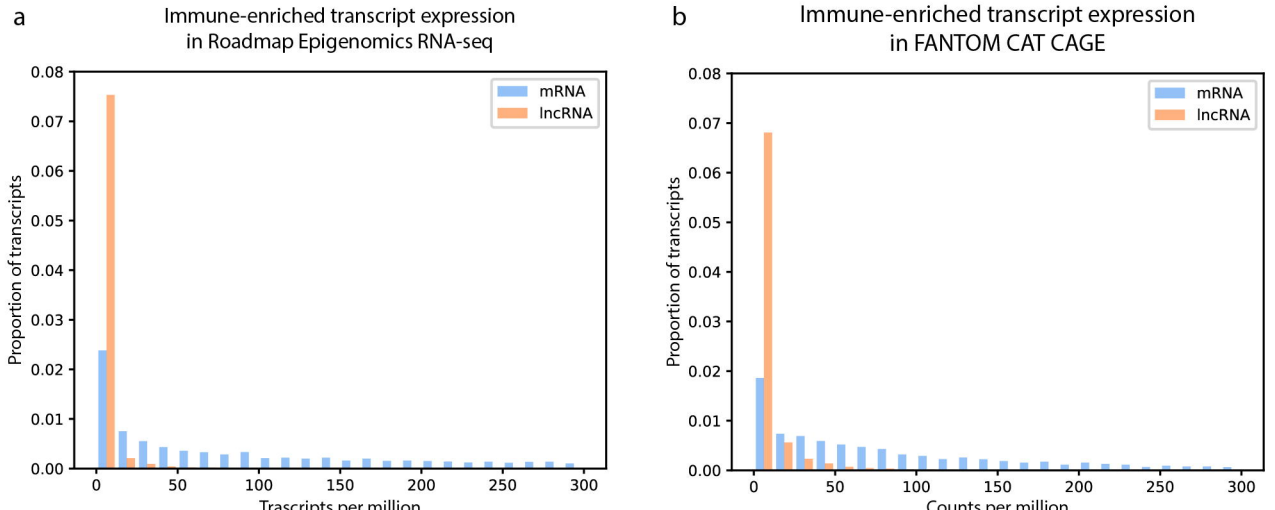
S2 FIG. ENRICHMENT OF IMMUNE-RELEVANT ENHANCERS AMONGST RA SUSCEPTIBILITY VARIANTS AFTER CONDITIONING ON IMMUNE-ENRICHED TRANSCRIPTS

The influence of FANTOM CAT immune-enriched lncRNA and mRNA was fixed in a probabilistic model of RA susceptibility to confirm the independent enrichment of immune-relevant enhancer chromatin states.


Enhancers
Transcription start sites
Transcription
Repressed chromatin
IMR90 fetal lung fibroblasts Cell Line
H9 Cells
HUES6 Cells
HUES48 Cells
HUES64 Cells
H1 Cells
iPS-20b Cells
iPS-18 Cells
iPS DF 6.9 Cells
iPS DF 19.11 Cells
H1 Derived Neuronal Progenitor Cultured Cells
hESC Derived CD56+ Mesoderm Cultured Cells
hESC Derived CD56+ Ectoderm Cultured Cells
hESC Derived CD184+ Endoderm Cultured Cells
H1 BMP4 Derived Mesendoderm Cultured Cells
H1 BMP4 Derived Trophoblast Cultured Cells
H1 Derived Mesenchymal Stem Cells
Primary mononuclear cells from peripheral blood
Primary T cells from peripheral blood
Primary T cells effector/memory enriched from peripheral blood
Primary T regulatory cells from peripheral blood
Primary T helper cells from peripheral blood
Primary T helper naive cells from peripheral blood
Primary T helper cells PMA-I stimulated
Primary T helper 17 cells PMA-I stimulated
Primary T helper memory cells from peripheral blood 1
Primary T helper memory cells from peripheral blood 2
Primary T CD8+ memory cells from peripheral blood
Primary T helper naive cells from peripheral blood
Primary T CD8+ naive cells from peripheral blood
Primary monocytes from peripheral blood
Primary hematopoietic stem cells G-CSF-mobilized Female
Primary B cells from peripheral blood
Primary Natural Killer cells from peripheral blood
Bone Marrow Derived Cultured Mesenchymal Stem Cells
Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
Foreskin Fibroblast Primary Cells skin01
Foreskin Fibroblast Primary Cells skin02
Foreskin Melanocyte Primary Cells skin01
Foreskin Melanocyte Primary Cells skin03
Foreskin Keratinocyte Primary Cells skin03
Thymus
Fetal Thymus
Brain Hippocampus Middle
Brain Substantia Nigra
Brain Anterior Caudate
Brain Cingulate Gyrus
Brain Inferior Temporal Lobe
Brain Angular Gyrus
Brain_Dorsolateral_Prefrontal_Cortex
Adipose Nuclei
Psoas Muscle
Skeletal Muscle Female
Fetal Muscle Trunk
Fetal Muscle Leg
Right Atrium
Left Ventricle
Right Ventricle
Aorta
Duodenum Smooth Muscle
Colon Smooth Muscle
Rectal Smooth Muscle
Stomach Smooth Muscle
Fetal Stomach
Fetal Intestine Small
Fetal Intestine Large
Small Intestine
Sigmoid Colon
Colonic Mucosa
Rectal Mucosa Donor 29
Rectal Mucosa Donor 31
Esophagus
Gastric
Placenta Amnion
Ovary
Pancreatic Islets
Fetal Adrenal Gland
Placenta
Liver
Pancreas
Lung
Spleen
A549 EtOH 0.02pct Lung Carcinoma Cell Line
Dnd41 TCell Leukemia Cell Line
GM12878 Lymphoblastoid Cells
HeLa-S3 Cervical Carcinoma Cell Line
HepG2 Hepatocellular Carcinoma Cell Line
HMEC Mammary Epithelial Primary Cells
HSMM Skeletal Muscle Myoblasts Cells
HSMM cell derived Skeletal Muscle Myotubes Cells
HUVEC Umbilical Vein Endothelial Primary Cells
K562 Leukemia Cells
Monocytes-CD14+ RO01746 Primary Cells
NH-A Astrocytes Primary Cells
NHDF-Ad Adult Dermal Fibroblast Primary Cells
NHEK-Epidermal Keratinocyte Primary Cells
NHLF Lung Fibroblast Primary Cells
Osteoblast Primary Cells
-4
0
4
log2 enrichment


a
lncipedia complete dataset
lncipedia high confidence dataset
MiTranscriptome lncRNA
FANTOM CAT lncRNA
FANTOM CAT mRNA
FANTOM CAT mRNA and lncRNA
b
FANTOM CAT immune enriched lncRNA
FANTOM CAT immune enriched mRNA
FANTOM CAT immune enriched mRNA and lncRNA
-4
0
4
log2 enrichment


Immune-relevant enhancers
FANTOM CAT immune-enriched lncRNA
FANTOM CAT immune-enriched mRNA
-4
0
4
log2 enrichment


a
Immune-enriched transcript expression
in Roadmap Epigenomics RNA-seq
mRNA
lncRNA
Proportion of transcripts
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
300
Trascripts per million
b
Immune-enriched transcript expression
in FANTOM CAT CAGE
mRNA
lncRNA
Proportion of transcripts
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
300
Counts per million