Reviewer: 1

Wang et al. presents the production and analysis of a large corpus of genomics data on the human brain as part of the PsychENCODE project. The analysis includes, in addition, brain data produced by other projects, creating likely the most comprehensive molecular resource of the human brain to date. I am a strong supporter of large-scale genomics projects. The resources that they generate serve as a foundation over which, through the work of many, knowledge is built about biological phenomena—without such a foundation, building such a knowledge would be much more costly.

It is often difficult to identify a specific breakthrough in genomic projects, and sometimes, under pressure to highlight them, findings may be over-emphasized. This is not a mistake in Wang et al. The value of the genomics projects is in both the data that they produce, and the methods they pioneer to produce, and analyze the data, and to report the findings. In this regard, Wang et al. is, in my opinion, a very good large-scale functional genomics project. The authors produce substantial data, they enhance their data, by pooling it together with other data, they develop methods to look at data in novel ways, and they present and summarize their data and their findings through high-content graphical descriptions. Regarding specifically the last two points, I liked particularly their integrative deep-learning model (DSPN), through which the authors attempt to link genomic variants to organismic phenotypes (mostly brain disorders) through intermediate phenotypes. This will become increasingly relevant, as data on intermediate phenotypes (i.e biological imaging at different levels) will accumulate within the next years. Regarding the presentation of the data and findings I acknowledge the effort by the authors to create figures that synthesize a large amount of information; some of these figure require a lot of effort by the reader, and, in some cases, some additional guidance through the captions would be further acknowledge (see my specific comments, below). In addition, the paper includes some relevant resources and insights regarding cell composition, regulatory elements, eQTLs and GWAS.

As with all large genomic-scale projects, there is always some degree of arbitrarity in the angle that is chosen to present and analyze the data. For instance, the paper ignores comparisons with increasingly available primate brain genomic data, or, while the authors acknowledge that DSPN could be used to integrate neuroimaging data, no effort is made to integrate ample available neuroimaging data with associated genomics data. I am not suggesting that these analyses should be performed, maybe they are, but it is overall unclear to me what additional analysis of the data (these or others) are followed in more detail in companion papers (although some can be guessed from the references). What I find quite poor, compared to other genome projects is the Adult.PsychENCODE page. There are no query options, links to browsers that can be used by researchers interested in particular genes or genomic regions, no good data visualization tools, etc. It looks to me just a data repository. Also access to many data requires a synapse account.

I do not have strong issues with the paper itself; what follows is a number of comments/ suggestions:

1. I could not find information on the protocols employed for obtaining the data used by the consortium. What was the depth of the RNA-Seq or what genotyping platform was employed, etc?. I understand that data has been generated specifically for this project (according to Table S2.1), but this is not described in the paper, nor references to other papers describing this data are given.

2. Transcriptome Analysis. The authors employ a method to estimate cellular composition in bulk tissue RNA-Seq based on the expression signatures that they infer from single cell data. One important result is that a large fraction of the variance of gene expression across tissue samples can be explained by changes in the cellular composition of the samples. The authors estimate that this is about 85%. This has implications for analysis of differential gene expression, since it means, in particular that changes in gene expression detected between different conditions (cell types, tissues, species, etc.) may actually reveal changes in cellular composition. As a measure of “goodness” of their estimates of cellular proportions, the authors compute the R^2. Am I correct in assuming that in the way R^2 is computed in the paper, the underlying assumption is that expression variance is constant across all genes? If so, wouldn’t it make sense to provide the R^2 at the sample level (how well the expression for a single individual is explained with the estimated cell types, and see the distribution of these R^2 values, instead of calculating a global one).

Moreover, the authors used an alternative method, CIBERSORT, which apparently, it does not perform as well as their method (R^2=0.81, compared to 0.88 for their method. Although the wording is sort of confusing, because they say the variance is the R^2, which is actually the fraction of variance explained (also in the manuscript this number is 0.85)). Given the previous comments, maybe the authors could compute in addition the root square mean error (RMSE). In any case, I think that it would useful to know how correlated the two methods are (i.e. whether they tend to produce the same decomposition of not).

3. Related to the above, another interesting result of the paper is that the authors found changes in the cell fractions associated to phenotypes and disorders. However, it is not clear to me whether these changes are inferred over all brain samples, or over specific regions (i.e. only the Prefrontal Cortex). I believe that it would be of interest to investigate whether the changes occur similarly across different brain subregions.

4. Enhancers. I found surprising that the number of enhancers is much large in temporal cortex than in cerebellum (~43,000 compared to ~27,000). Any hypothesis about this?

5. Figure 3E. Is it not possible to distinguish the colors corresponding to the different tissues—even zooming in the figure. Also, what is the population of samples near (0,-2)?

6. I believe that the description of some figures and results could be improved. For instance, in Figure 3d it took me a while to understand what does “Transcriptome diversity” mean. I believe it simply means % of (100bp windows in) the genome transcriptionally active (although in the main text it appears to implicitly mean the number of genes transcribed); and “Inter sample transcriptome diversity” just the average transcriptome

diversity. Moreover, the caption of Figure 3d mentions triangles and circles, which do not appear in the figure.

7. QTL analysis. The authors claim a high replicability between their eQTLs and the GTEx brain eQTLs. To use this as support for the psychENCODE eQTLs, I believe that it would be of interest to show that this replicability is smaller when compared to eQTLs from GTEx tissues other than brain.

8. The authors use standard univariate approaches for isoform and cell fraction QTLs. These however are intrinsically multivariate phenotypes, in which the values of the different variables add to one (in the case of isoforms, if abundances are relative proportions). Testing them independently ignores the strong dependency between them. There are now a number of methods that test associations with multivariate phenotypes, which are probably more appropriate in these cases.

9. I am not sure I understand how the cell-type differences are factored out to identify 200,729 trans-eQTLs which “represent variant-expression associations largely unexplained by changing proportions of cell types”. I do not think this is explained in the supplementary information. Also, why not a similar approach (i.e. using the cell fractions as co-variates) to identify cis-eQTLs truly associated to changes in gene expression and not to cell fractions?

10. Regulatory Networks. Figure S6.1 The caption is confusing. I think that there is no description for panel C and panel D is missing. Also in Figure S6.2, maybe it is obvious, but what do CP and GZ stand for? In Figure 5F. what are the orange dots?

11. Integrative deep-learning model. Figure 6D in general shows that DSPN improves phenothype prediction over other simpler methods, such as logistic regression, that do not take into account intermediate phenotypes. However, this is not the case for ethnicity, in which DSPN performs worse than using only genotype information. Obviously, ethnicity is fully determined by genotype, and adding additional information may mask the genotype signal. Maybe the authors could comment under which circumstances DSPN may not be the optimal approach. Also, the figure is an example of the need of some better explanation. As employed by the authors DSPN predicts binary phenotypes. But in Fig 6D one of the phenotypes investigated is age. It is not obvious from the caption of the text that age is being binarized

Reviewer: 2

In this manuscript, the authors describe an impressive wealth of data generated within the PsychENCODE consortium, including genotype, transcriptome, chromatin (including Hi-C) and single cell data from over 1800 individuals. These data were then integrated with existing functional genomic resources to build a unique resource for functional genomics in brain research. The authors not only catalogue the findings but also underline the relevance of this resource by presenting more extended functional annotation of schizophrenia GWAS hits and by using these data to develop a deep learning model that integrate these molecular layers with genotypes and increases disease prediction.

Overall, from the wealth of data, the authors present more detail on relating single cell gene expression data to bulk RNA seq. This analysis, especially with the power of large sample sizes is highly relevant for the community.

The second focus is on enhancers. Here there is a limitation as only two chromatin marks are investigated in depth. This should be presented as a limitation. This also introduces a limitation into the transcriptome epigenome comparison.

The QTL analyses are very detailed and present for the first time QTL ranging from expression, chromatin, splicing-isoform, and cell-fraction QTLs, with interesting intersection presented. As for the cell fraction QTLs, this is very novel, but more detail on this analysis needs to be given to better understand how the SNPs drive this specific QTL.

The data presented so far are then integrated into regulatory network, investigating dual to multi-QTL. Here some more detail should be given on the mentioned full regulatory network beyond the number of linkages. These networks are then linked to GWAS data, showing strong enrichment for brain disorders and highlighting new candidate genes for schizophrenia. This important analysis seems somewhat buried in the supplements and this reviewer could not really find a link to the list of these genes. Also in this section CACNA1C is mentioned as an example but CHRNA2 is then shown in the figure (5F).

Finally, the authors present an integrative deep learning model and apply it to differentiate between cases and controls using multilayer molecular data. Again it would be helpful if the authors could briefly explain in the main text what kind of samples and data were used to test the full models vs. just genotype alone.

Overall, the authors present data and analyses from a unique resource leading to novel insights about regulatory elements in brain and their relationship to psychiatric disease. The depth of analysis – from bulk sequencing to single cell with multiple layers of molecular information is unique, especially paired with the large sample size and the work definitely deserves publication in a journal for a broad audience such as Science. Overall, however, the paper is extremely dense and a lot of very important information is now only in the supplement. Here additional brief summary sentences in the main text may help to enhance the link between the main text and the supplements and could avoid that readers miss some important results.

Most importantly, the authors have made their data and analyses available to the public in a very well designed and intuitive website, so that this resource can be easily accessed by other researchers.

Reviewer: 3

The manuscript from Wang et al, entitle “comprehensive functional genomic resource and integrative model for the adult brain” provides for a comprehensive presentation of new genomic resources from the PsychENCODE consortium, including ~5,500 datasets from ~1900 individuals. Through a series of analyses, the authors assess i- the source of inter-sample heterogeneity in expression profiles, arguing for changes in proportion of cell types accounting for over 85% of the variation, ii- define the cis-regulatory landscape and iii- report QTLs for expression, chromatin, splicing and cell-type-proportion changes in the reference brain. Finally, taking all the data into account, the authors present a new method to decipher disease predisposition and extend the list of genes targeted by GWAS hits for psychiatric disorders.

Overall, this manuscript provides a comprehensive analysis of the PsychENCODE data, addressing the primary needs in analyzing such large genomic datasets. It also provides biological insight into population heterogeneity, exploiting single-cell RNA-seq to capture the contribution of changes in the proportion of cell types between samples to justify inter-sample variation. It also addresses the functional characterization of genetic risk to psychiatric diseases.

Major concerns:

1. This manuscript is rich in information and would benefit from ensuring that all critical information be presented in the main text. Currently the reader has to rely heavily on the supplementary section to adequately understand the work. In addition, the following should be considered:

1.1 The type and number of samples available for each dataset should be clearly mentioned especially when multiple resources have been combined together for different analysis.

1.2. Every section should mention the exact supplementary section being referred to instead of just by ref(15).

1.3 All supplementary figures should be referenced in the text.

2. The source code of the deep-learning model should be made available.

3. Legend to figure 3 needs to be revised. For instance, “Panels E and F are drown similarly to D” should read “similarly to B”. Also, Panel D need to be explained before E and F. Also, the figure states “(D) … for coding (circle) and noncoding (triangle)…” while only diamonds are presented in the figure. The legend should also include a description of the meaning for the arrows.

4. The DSPN model uses a conditional deep boltzmann machine which requires binarized representations of the data used as also stated by the authors. It should be explained how gene expression and enhancer activity were binarized to fit the model.

Minor comments

5. Recommend a quantitative as opposed to qualitative report of the results in the text. For instance, on page 6 “…the proportions of excitatory neuron Ex4 and Ex5 were associated with the most”,

this last word should be replaced by a quantitative assessment. (also consider completing that sentence with “fQTLs”).

6. Figure 5C, the number of data points should be clearly mentioned for each boxplot on the figure.

7. Figure 4B, consider adding the genotype next to the H3K27ac signal on the figure

8. Change “Epigenomics Roadmap” on page 3 first paragraph to “Roadmap Epigenomics Project (Roadmap)” and maintain the use of “Roadmap” in subsequent sections

9. The resolution on Figure 1, 3, 4 and 5 was poor. Please provide better quality figures.

10. Supplementary figure 3.2, page 22, y-axis add “b” in “number”

Reviewer: 4

The authors use bulk and single cell RNAseq, H3K27ac, and Hi-C data to explore the molecular structure of the adult human brain.

While it is clear that the raw and "pipeline processed" data from these studies is available through the PsychENCODE portal, it is my strong opinion that the more highly processed data elements used to detail individual findings in this report (and more generally in all large data resource annoucements of this type) should be made more easily accessible to readers. Each time an analysis is done and a finding reported, the processed data involved should be easily and directly accessible to any reader. In particular, the cell type data derived from the single cell data, the list of 80K enhancers in the reference brain as well as the 120K enhancers in the larger collection, the various QTLs identified, the Hi-C maps and regulatory networks, genes linked to GWAS variants - all of these processed data elements can be useful tools in additional in silico analysis and should be at the finger tips of readers.

Perhaps particular attention can be paid to the deep learning model in this manner. The authors stress its interpretability. The structure, utility and exact location (for download) of input, intermediate, and output elements of this approach should be made clear to enable such secondary use of this analysis.

In the "Linking GWAS variants to genes" section, it should be noted that the linking of GWAS variants to genes in this report uses only data from the adult brain, while much of risk for schizophrenia, ASD and bipolar disorder has been mapped to fetal development. Therefore, this analysis may be missing a large proportion of the relevant expression and epigentic information to make these links.

A minor point: the NMF algorithm used is not noted in either the main text or methods - what algorithm was used and does it produce stable/consistent results across multiple runs?

Reviewer: 4