Supplemental File 1 **Predicted genes** Gene predictions in GTF format against reference genome *human_g1k*. Transcriptome-based predictions are marked with *Cufflinks*; *ab initio* predictions are marked with *FGENESH*.

Supplemental Text 1

Human specimens. The description of postmortem brain specimens is presented in **Supplemental Table S1**.

Genomic DNA library preparation, whole genome sequencing and analysis. Four brain samples were used for whole genome sequencing: A5, A9, S14 and C13. 2 ug of genomic DNA was sheared on Covaris S2 system (Applied Biosystems). Genomic libraries were prepared according to Paired-End DNA Sample Prep Kit protocol (Illumina). 101 bp PE sequencing was performed on Illumina's HiSeq 2000 achieving average 35-fold genome coverage (specifically, 36-fold A5 and S14, 22-fold for A9 and , 24-fold C13). PE reads were mapped to GRCh37 reference genome with BWA program (1). Duplicate reads were marked with Picard package. SNP calling was performed according to GATK best practice variant calling protocol v3 (2). All stages of genome analysis were compiled into our whole genome sequencing analysis pipeline (http://rogaevlab.ru/ngs-pipeline).

Data visualization. We have routinely used IGV (3, 4) for data visualization and figure preparation; we have also used R and, in particular, ggplot2 (5) package for plotting.

ChIP-seq data analysis. ***Read mapping.*** Initial set contained 64 H3K4me3 histone methylation ChIP-seq samples, from which 60 were from neuronal and 4 from nonneuronal brain tissues. We mapped all samples to reference human genome GRCh37 using Bowtie 2.0.2 (6) with default parameters. Rate of unmapped reads varied from 0.8% to about 50.7% with non-uniquely mapped reads varied from 9.4% to 26.5%, respectively Then, we removed potential PCR-duplicates and non-uniquely mapped reads (MAPQ = 0). We additionally excluded 3 neuronal samples with less than 3 million reads mapped. Peaks were called for all samples using MACS 1.4.2 (7). We ran it with P-value thresholds $1e^{-5}$ and $1e^{-10}$ against input, but used $1e^{-10}$ in downstream analysis. We also excluded all peaks with length less than 200bp. Promoter coverage was calculated for each sample using BEDTools (8). We defined promoters as regions of 2 kb upstream and 2 kb downstream from TSS of each gene. All such regions were merged using BEDTools and total promoter coverage was calculated for all samples. We used these values for normalization. Base coverage for each position in peak were totaled using an in-house tools. These values were multiplied to 10^7 and divided to total promoter coverage of corresponding sample and used as measures of H3K4me3 occupancy for most of downstream analyses. ***Quality control.*** We excluded three neuronal samples from further analysis, due to shallow sequencing coverage (< 2 million reads mapped). As anticipated, we found H3K4me3 peaks in genes on Y-chromosome in all males only and in *XIST* gene, which is essential for inactivation of one X-chromosome homolog (9) exclusively in females (**Supplemental Fig. S1**). Furthermore, we assessed the purity of neuronal chromatin for each sample by scanning genes known to be expressed in neurons but not in glia and vice versa (10-13). We found that all neuronal samples showed robust specific H3K4me3 peaks exclusively for neuronal genes and lacked peaks for glial-specific genes. Correspondingly, 3 of 4 non-neuronal samples showed signals for glial-specific genes with no peaks for neuronal-specific genes (the fourth sample showing also neuronal signals was excluded from further analyses) (**Supplemental Fig. S1a**). As expected, H3K4me3 states detectable by ChIP-seq showed highly specific locus-based signatures in brain cells (**Supplemental Fig. S1**).

RNA-seq data analysis. RNA-seq data sets for the four total cortical specimens and 20 samples from grey and 20 samples from white matter specimens were aligned to genome GRCh37 reference with Tophat v2.0.4 program (14) with supplying Ensembl gene models v. 69. Transcripts were assembled for each library separately with Cufflinks software (15). Additionally, we used 18 published RNAseq datasets from superior temporal gyrus (ERP001304 project) (16), nine of which were patients with SZ and nine were controls. The same analysis as above was performed. To detect new transcripts missing in Ensembl and UCSC databases, we merged gene models produced by cufflinks for 40 RNAseq libraries as well as gene models from 18 libraries of ERP001304 project. After comparison to Ensembl v87, v91 and UCSC genome browser we've identified 1526 new potential genes.

Ab initio gene predictions. To predict genes using only reference genomes sequence in close proximity to H3K4me3 peaks, we selected 200 kb windows in both direction to the peak and performed gene predictions using FGENESH (17) and selected only predictions with TSS inside the peak.

Coding potential of previously undescribed genes. To search for novel transcripts mapped to neuronal gene promoters marked by H3K4me3, we used transcripts assembled from the RNA-seq for 58 brain specimens (superior temporal gyrus, grey and white matters) as described above. To evaluate coding potential of previously undescribed genes, we performed the following procedures: we generated cDNA sequences (CDS) for each transcript with gffread utility script from Cufflinks (15). For transcripts with unknown direction, both strands we used to generate cDNA sequences. All sequences were analyzed with Coding Potential Calculator (18) to predict the ORF for each transcript and to produce a score for potential coding or non-coding (above or below 0, respectively). We classify a gene to be potentially protein-coding if any of transcripts have a coding potential greater than 1. Coding Potential Calculator provides peptide sequences for predicted ORFs. We identified protein domains in these sequences by searching PFAM (19) database with HMMER (20).

Analysis of correlation between genetic CNVs and epigenetic-H3K4me3 variations. To identify putative CNVs, we run both FREEC (21) and CNVnator (22) with default parameters (using 1000bp bins for CNVnator). From CNVnator output, we also excluded loci with q0 > 0.5. Given that different algorithms for CNV discovery deviate from each other (23), we focused our analysis only on high-confidence CNVs, that were identified by both of these tools. To study the interactions of CNV and H3K4me3 peak size, we compared normalized tag densities for four individuals in peaks, that overlap CNV regions. We considered only peaks where at least one individual has a CNV. We evaluated the average peaks size for individuals without CNV (copy number = 2) and quantified fold change for all four individuals (there can be two or three individuals without CNV, in these cases fold change is not equal to 1; if there is only one individual without CNV, fold change for him is exactly one). Next, we separated all fold changes into three groups: a) where copy number < 2 (deletions) b) where copy number == 2 and c) where copy number > 2 (duplications) and quantified the difference between these groups with Mann-Whitney-Wilcox non-parametric test. We also estimated the correlation between the copy number and fold change across peaks and individuals. To increase robustness, we excluded peaks with elevated rate of reads with low mapping quality (75% of reads have MAPQ < 10) because CNV calls are based on longer reads (2x100bp vs 36bp).

Effect of SNVs and INDELs on individual peak induction in four individuals. To select individual-specific variants in the four individuals, we first selected variants within any of 29,547 peaks using bcftools and annotated singleton variants with an in-house tool, based on bio-vcf package (24). Annotation of variants for dbSNP presence and population frequency was performed with Variant Effect Predictor (25), evolutionary conservation was based on PhyloP46 scores (26) from UCSC genome browser (27) (PhyloP46 > 1.0 for a conservative site) and effect on CpG sites was evaluated with an in-house tool based on rust-bio and rust-htslib packages (28). We evaluated the significance of overlap using hypergeometric test without correcting for multiple testing.

NUP210L genotyping, allele imbalance and expression analysis. Of five individuals showing neuronal H3K4me3 peak in NULP210L locus, we were able to test rs114697636 for four individuals by sequencing or direct inspection of ChIP-seq reads. The rare rs114697636 G-allele was identified in all four subjects. Genotyping of rs114697636 at 5'-region of NUPL210L of brain specimens was performed by Sanger sequencing of PCR product amplified from genomic DNA using the following primers: NT_F-CGCCGGACAGCCAGTCAT; NT_R-GCTGACGTCACGCCTGTG. Genotyping of rs114697636, rs11264875 at 32nd exon of NUPL210L in brain specimens was performed from genomic DNA and cDNA. For amplification from genomic DNA, the following oligonucleotide primer sequences were used: F875-GCCCAGGCTGTGAACAGAGGG; R875-TGGCTGGCTTCTAATTGGCTGGC. For target region amplification from cDNA, the following two primer oligonucleotides were used: F-357-CCACACACACAATACCCAGC; R-357-AAAATCATTGCAGTCCCCGG. Sample C28 cDNA was additionally amplified with the following primers: F-TTGCTGCATATTGGACCAGG; R-ACTGAAATCTGAATGGACCTGA.

Transcription factor binding site prediction. Transcription factor binding sites were predicted with PERFECTOS-APE (29) using HOCOMOCO (30) and JASPAR (31) motif databases using default parameters.

Gel shift assay for NUP210L. ***Preparation of nuclear extracts.*** Cells were washed with ice-cold PBS and then pelleted by centrifugation at 300 g for 2 min. The cell pellet was resuspended in 1 ml of Buffer 1 (10 mM HEPES, pH 7.9, 10 mM KCl, 1 mM dithiothreitol, 0.5 mM spermidine, 0.15 mM spermine, 0.1 mM EDTA, 0.1 mM EGTA, 0.5 mM PMSF, 16 Halt protease inhibitor cocktail (Thermo Scientific)) and placed on ice to swell for 15 min. After addition of 62 ml of 10% (w/v) Nonidet P-40, the sample was gently vortexed for 10 s and then centrifuged at 400 g for 5 min at 4°C. The nuclear pellet was resuspended in 100 ml of Buffer 2 (20 mM HEPES, pH 7.9, 25% (w/v) glycerol, 420 mM NaCl, 1.5 mM MgCl2, 0.2 mM EDTA, 1 mM dithiothreitol, Halt protease inhibitor cocktail (Thermo Scientific)) and incubated for 20 min on ice followed by a centrifugation at 10,000 g for 10 min at 4°C. The supernatant containing the nuclear proteins was stored at -70°C.

Electrophoretic mobility shift assays. For allelic version analysis, double stranded oligonucleotides V1-GGCTGTAGTTCAGCGGGAACCC and V2-GGCTGTAGTTGAGCGGGAACCC were synthesized with 5'-CAGT tetranucleotide overhangs and SNP position in the center. Introduction of labels in the DNA probe was conducted using fill in of shortened 3' ends with Klenow fragments of DNA polymerase I. The reaction was conducted for 5 min at room temperature in 10 µl of reaction mixture which contained 0.01 nmol oligonucleotides, 1 µl of 10x buffer for labeling (500 mM Tris-HCl, pH 8.0, 100 mM NaCl, 100 mM $MgCl_2$, 1 mM DTT, 2 mM dGTP, 2 mM dTTP, 2 mM dCTP), 2 active units of Klenow fragment, and 10 mCi (a-^{32}P)dATP. The protein nuclear extract was diluted to the desired concentration of salt with dilution buffer (20 mM HEPES, pH 7.9, 25% (w/v) glycerol, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1 mM dithiothreitol, Halt protease inhibitor cocktail (Thermo Scientific)) and incubated with sheared salmon sperm DNA (100 ng of DNA per 7 mg of total protein) for 10 min at 0°C. After that 4 mg of extract was added to the probes which contained 100 fmol of radioactive labeled oligonucleotide. For competition experiments, cold double-stranded oligos V1, V2 and TAT_GR-cagtTGCTGTACAGGATGTTCTAGC (corresponding GR binding site from tyrosine aminotransferase gene) were added to the reaction mix at 0.1 pmol and incubated on ice for 20 minutes before addition of the labeled probe. After incubation at room temperature for 15 min the mixture was subjected to electrophoresis in 4.5% PAAG in 0.56TBE (89 mM Tris-borate, 89 mM H3BO3, 2 mM EDTA at 4°C). The gel was exposed to X-ray film.

References

1. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760

2. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303

3. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011) Integrative genomics viewer. *Nat Biotech* **29**, 24–26

4. Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192

5. Wickham, H. (2010) Ggplot2: Elegant Graphics for Data Analysis. Springer, Dordrecht; New York

6. Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**, 357–359

7. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137

8. Quinlan, A. R. and Hall, I. M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842

9. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., and Brockdorff, N. (1996) Requirement for Xist in X chromosome inactivation. *Nature* **379**, 131–137

10. Zuccato, C., Ciammola, A., Rigamonti, D., Leavitt, B. R., Goffredo, D., Conti, L., MacDonald, M. E., Friedlander, R. M., Silani, V., Hayden, M. R., Timmusk, T., Sipione, S., and Cattaneo, E. (2001) Loss of huntingtin-mediated BDNF gene transcription in Huntington's disease. *Science* **293**, 493–498

11. Pham-Dinh, D., Mattei, M. G., Nussbaum, J. L., Roussel, G., Pontarotti, P., Roeckel, N., Mather, I. H., Artzt, K., Lindahl, K. F., and Dautigny, A. (1993) Myelin/oligodendrocyte glycoprotein is a member of a subset of the immunoglobulin superfamily encoded within the major histocompatibility complex. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7990–7994

12. Carlock, L., Vo, T., Lorincz, M., Walker, P. D., Bessert, D., Wisniewski, D., and Dunbar, J. C. (1996) Variable subcellular localization of a neuron-specific protein during NTera 2 differentiation into post-mitotic human neurons. *Brain Res. Mol. Brain Res.* **42**, 202–212

13. Snipes, G. J., Suter, U., Welcher, A. A., and Shooter, E. M. (1992) Characterization of a novel peripheral nervous system myelin protein (PMP-22/SR13). *J. Cell Biol.* **117**, 225–238

14. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013) TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36

15. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* **28**, 511–515

16. Wu, J. Q., Wang, X., Beveridge, N. J., Tooney, P. A., Scott, R. J., Carr, V. J., and Cairns, M. J. (2012) Transcriptome sequencing revealed significant alteration of cortical promoter usage and splicing in schizophrenia. *PLoS ONE* **7**, e36351

17. Salamov, A. A. and Solovyev, V. V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10**, 516–522

18. Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G. (2007) CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**, W345–W349

19. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014) Pfam: The protein families database. *Nucl. Acids Res.* **42**, D222–D230

20. HMMER (http://hmmer.org/).

21. Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2012) Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425

22. Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984

23. Duan, J., Zhang, J.-G., Deng, H.-W., and Wang, Y.-P. (2013) Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS ONE* **8**, e59128

24. Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J., and Katayama, T. (2010) BioRuby: Bioinformatics software for the Ruby programming language. *Bioinformatics* **26**, 2617–2619

25. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070

26. Pollard, K., Hubisz, M., and Siepel, A. (2009) Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res.* gr.097857.109

27. Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., and Kent, W. J. (2015) The UCSC Genome Browser database: 2015 update. *Nucl. Acids Res.* **43**, D670–D681

28. Köster, J. (2016) Rust-Bio: A fast and safe bioinformatics library. *Bioinformatics* **32**, 444–446

29. Vorontsov, I., Kulakovskiy, I., Khimulya, G., Nikolaeva, D., and Makeev, V. (2015) PERFECTOS-APE–predicting regulatory functional effect of SNPs by approximate P-value estimation.

30. Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Soboleva, A. V., Kasianov, A. S., Ashoor, H., Ba-Alawi, W., Bajic, V. B., Medvedeva, Y. A., Kolpakov, F. A., and Makeev, V. J. (2016) HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.* **44**, D116–125

31. Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014) JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucl. Acids Res.* **42**, D142–D147