Science. Author manuscript; available in PMC 2010 May 11.

Published in final edited form as:
Science. 2009 September 4; 325(5945): 1246–1250. doi:10.1126/science.1174148.

Common regulatory variation impacts gene expression in a cell type dependent manner

Antigone S. Dimas[1,#], **Samuel Deutsch**[2,#], **Barbara E. Stranger**[1,3,#], **Stephen B. Montgomery**[1,#], **Christelle Borel**[2], **Homa Attar-Cohen**[2], **Catherine Ingle**[1], **Claude Beazley**[1], **Maria Gutierrez Arcelus**[1], **Magdalena Sekowska**[1], **Marilyne Gagnebin**[2], **James Nisbett**[1], **Panos Deloukas**[1], **Emmanouil T. Dermitzakis**[1,4,*], and **Stylianos E. Antonarakis**[2,*]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, CB10 1HH, Cambridge, UK
[2]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, CH-1211, Switzerland [3]Division of Genetics, Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, 02115 USA

Abstract

Studies correlating genetic variation to gene expression facilitate the interpretation of common human phenotypes and disease. As functional variants may be operating in a tissue-dependent manner, we performed gene expression profiling and association with genetic variants (SNPs) on three cell types of 75 individuals. We detected cell type-specific genetic effects, with 69 - 80% of regulatory variants operating in a cell type-specific manner and identified multiple eQTLs per gene, unique or shared among cell types and positively correlated with the number of transcripts per gene. Cell type specific eQTLs were found at larger distances from genes and lower effect size similar to known enhancers. These data suggest that the complete regulatory variant repertoire can only be uncovered in the context of cell type specificity.

Variation influencing gene expression can manifest itself as gene expression differences among populations, among individuals in a population, among tissues, and in response to environmental factors. The genetic basis of the first two types of gene expression variation has been investigated with the quantification of mRNA in one tissue, and the identification of genetic variants correlated with expression variation (eQTLs) in a single or multiple populations (1-7). The complexity in higher eukaryotes however results in a vast set of highly specialized cell types and tissues. Some genes exhibit ubiquitous patterns of expression while others display tissue-specific activity (8-10). Although some studies have identified eQTLs in human (11-13) and mammalian tissues (14, 15), a systematic study comparing eQTLs across different cell types while controlling for confounding associations (population samples, differences in technology or statistical methodology) is lacking. Documenting tissue-specific genetic control of gene expression variation may connect cellular activities in health and disease (11, 13, 16). Efforts to use genomic information to interpret the biological effects of such variants are hindered by the limited availability of the relevant human tissues.

We investigated 85 individuals of the GenCord project (a collection of cell lines from umbilical cords of individuals of Western European origin) to identify cis eQTLs involved

*correspondence should be addressed to ETD (emmanouil.dermitzakis@unige.ch) or SEA (stylianos.antonarakis@unige.ch).
#these authors contributed equally
[4]Present address: Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, CH-1211, Switzerland

Europe PMC Funders Author Manuscripts

in gene expression variation in three cell types: primary fibroblasts, EBV-immortalized B-cells (lymphoblastoid cell lines or LCLs) and T-cells. Umbilical cord was chosen because it is readily available and allows the acquisition of multiple cell types. Sample collection was performed on full term or near full term pregnancies to ensure homogeneity for sample age.

mRNA levels were quantified in primary fibroblasts, LCLs, and primary T-cells for 48,804 probes with the illumina WG-6 v3 Expression array. We analyzed data from 22,651 probes uniquely mapping to 17,945 autosomal RefSeq genes (15,596 Ensembl genes). Samples were genotyped on the illumina 550K SNP array. Following quality control (SNPs with missing data were removed) and minor allele frequency filter (MAF ≥ 5%), 394,651 SNPs were used for association testing. Principal components analysis (PCA) detected 10 potential outlier individuals from the genotype data (17) (Fig. S1) and were subsequently removed from the analysis. eQTL discovery and all other properties of the results for 75 vs. 85 individuals were almost identical (Fig. S2).

We explored associations in *cis*, by testing all SNPs within a 2Mb window centered on the transcription start site (TSS) of a gene. Spearman rank correlation (SRC) tested for association between SNP genotype and mRNA levels (intensities), after normalization and transformation. SRC performs similarly to linear regression (7), but is not sensitive to outliers in the gene expression data, which reduces power. A total of 6,083,130 tests were performed and significance thresholds for each gene were assigned through 10,000 permutations of expression values, as described (6, 7, 18). For 75 individuals at the 0.001 Permutation Threshold (PT), we discovered 427, 442 and 430 genes with significant *cis* associations in fibroblasts, LCLs and T-cells respectively, with an estimated false discovery rate (FDR) of 4% (Fig. S2A and B; Tables S1 and S2). For the less stringent permutation threshold of 0.01 we discovered 2,146, 2,155 and 2,046 genes with an estimated FDR of 7% (Fig. 1 and SOM) (19). The range of allelic effects was estimated by comparing the median expression in the two homozygote classes for the eQTL SNPs and the 95% confidence limits of fold change were between 1.07 and 2.65 (Fig. S7).

We assessed whether we can replicate the LCL eQTLs from previous studies. eQTLs from the CEU HapMap LCLs (7, 17) were replicated in the GenCord LCLs. Both populations are of European descent and share similar allele frequency spectra. Due to the differences in probe sequence content between the illumina v1 array (used for HapMap CEU) and the v3 array (used here) we could only compare a small subset of those SNP-probe associations. Of the 5,898 SNP-probe pairs surviving the 0.001 PT in HapMap CEU, 137 SNP-probe pairs (44 probes, some associated with multiple SNPs) were also tested in GenCord LCLs. Of the 137 SNP-probe tests 114 had p-values <0.001 (83%) (Fig. S3). Therefore previously detected eQTLs were well-replicated, despite the long separation time between tests of these cell lines, demonstrating the stability of transformed B-cells.

We interrogated the cell type specificity of regulatory effects by exploring genes with *cis*-eQTLs that were: a) shared in all three cell types, b) shared in two cell types, and c) cell type-specific. At the 0.001 PT, we identified a non-redundant set of 1,007 genes with *cis*-eQTLs of which 86 (8.5%) were shared among all three cell types, 120 (12%) shared in two of the cell types, and 801 (79.5%) were cell type-specific (Tables S1, S2). The proportion of cell type-specific eQTLs was similar to previous estimates of eQTL tissue specificity and alternative splicing reported in a study interrogating two tissue types sampled however from different individuals (20).

The degree of eQTL sharing across cell types (Fig. S4) is overestimated because the eQTLs for the same gene are not necessarily identical genetic variants (see below). Of the genes with *cis*-eQTLs common to two or more cell types, 124 (12.3%) were shared between

fibroblasts and LCLs, 121 (12.0%) were shared between fibroblasts and T-cells, and 133 (13.2%) were shared between LCLs and T-cells. Increased eQTL sharing between LCLs and T-cells is most likely due to the similarity of these cell types. We observed a striking prominence of tissue specificity with 268 (26.6% of total), 271 (26.9%), and 262 (26.0%) of gene eQTL associations found only in fibroblasts, LCLs and T-cells respectively.

To test if the eQTL cell type specificity arises from differential expression between cell types we compared medians and variance of gene expression. We found that genes with cell type-specific eQTLs had significantly higher expression variance in the eQTL cell type (Mann-Whitney U, $P < 0.0001$ for all comparisons). Medians for the same genes were either marginally significantly or not different, meaning that genes included in this analysis were largely expressed in all cell types. This suggests that the majority of cell type specificity is not a result of differential gene expression levels between cell types, but due to cell type-specific use of regulatory elements.

To dissect the overlap of *cis*-eQTLs across cell types, we compared the direction of the allelic effect for eQTLs significant in two or more cell types. The direction (sign of Spearman rho) was in complete agreement for all pairwise cell type comparisons at the 0.001 PT (Fig. S5). Thus regulatory variants are active across cell-types in the same manner. To assess the strength of cell type specificity, we performed repeated-measures ANOVA (RMA). Cell type specificity was reflected in the SNP x cell type interaction term (the part of the equation that tests SNPs effects dependent on the cell type), where cell type-specific associations are expected to be significant. We found 61% enrichment of low p-values in cell type-specific eQTLs, (quantified by estimation of FDR (21)) (Fig. S6). No enrichment was observed for cell type-shared eQTLs. RMA however is limited as the power to detect an interaction term is never maximized due to lack of allelic effect reversal between cell types.

We further used allele specific expression assays to validate a subset of cell type-specific eQTLs. We measured the ratio of the two alleles of transcript SNP in RNA samples of individuals who were double heterozygotes for both the eQTL and the transcript SNP. For 35 transcript SNPs (7 from fibroblasts, 14 from LCLs, and 14 from T-cells) we observed extensive allelic imbalance (ratio of the abundance of the transcripts of the two alleles) for the eQTL cell type; this was highly significantly different from the ratios of the same SNPs in cell types without the eQTL (paired t-test. $P = 5.6 \times 10^{-7}$) (Fig. S7). Therefore, the eQTL cell type specificity has been experimentally confirmed.

Shared associations among cell type increased slightly at relaxed significance thresholds for one cell type, due to the so called “winner’s curse” (22-24) (Fig. S8). This states that the effect sizes discovered when applying specific statistical significance thresholds are inflated compared with the true effect size. Consequently, the discovery sample usually achieves higher significance than replication samples. Even with relaxed thresholds however, over half of the associations we detected remain cell type-specific. We selected significant SNP-probe pairs from one cell type, and explored their nominal (uncorrected) p-value distribution in the other two cell types. These distributions were enriched for low p-values, reflecting those associations that are shared between cell types (Fig. 2). When we removed SNP-probe associations with significant associations in the secondary cell type (i.e. shared associations at the same and lower significance threshold) the resulting nominal p-value distributions demonstrated only small enrichment for low p-values. We quantified the fraction of significant *cis* eQTLs from one cell type that is not nominally significant (p-value prior to correction > 0.05) in either of the other two cell types. We estimate that 54%, 50% and 54% of *cis*-eQTLs in fibroblasts, LCLs and T-cells respectively, are cell type-specific, amounting to 69% of all *cis*- eQTLs at 0.001 PT Therefore the limited overlap of *cis*-eQTLs between

cell types is unlikely to result from winner's curse and supports the conclusion that a substantial fraction of eQTLs are cell type-specific.

As previously observed (7, 25) we found that the strength and density of *cis* eQTLs decay symmetrically with increasing distance from the corresponding gene's TSS (Fig. 3A). To better understand the independent regulatory effects we mapped eQTLs into recombination hotspot intervals and subsequently further controlled for linkage disequilibrium (see SOM) (26). At the 0.001 PT, we observed that 5.1% of associated genes possess more than one independent interval carrying an eQTL (Table S3). To further dissect the regulatory variant sharing between genes, we compared the overlap of independent eQTLs (i.e. regulatory intervals rather than genes) across cell types. When all intervals with a 0.001 PT eQTL were considered, only 6.9% were found to be shared across all three cell types. 9.7% were shared in two cell types and 83.4% were cell type-specific (Fig. 4A, Table S4). The degree of overlap increased as independent eQTLs for genes with shared expression associations across cell types were analyzed (Fig. 4B, C). In all cases however, a substantial fraction of independent eQTLs were cell type-specific.

Cell type-shared eQTLs tend to have larger effects (Spearman correlation coefficient rho), higher significance, and cluster tightly around the TSS (Figs. 3B and S9). Cell type-specific eQTLs, however, have lower effect sizes and are more widely distributed around the TSS (Fig. 3C). This is in agreement with the finding that enhancer elements, which tend to be found at greater distances from the gene, show greater tissue-specificity than basic regulatory elements (27). The number of eQTLs per gene was significantly correlated with number of transcripts per gene (Pearson's correlation coefficient = 0.049, p-value= 0.117 at 0.001 PT and Pearson's correlation coefficient = 0.105, p-value<0.0001 at 0.01 PT). This suggests that regulatory complexity correlates with transcript complexity. Single eQTLs SNPs were also found to influence expression of multiple genes. At the 0.001 (and 0.01) PT, over 6% (19%) of eQTL SNPs were associated with the expression of more than one gene (Fig. S10).

We used Gene Ontology (GO) (28) to compare the properties of cell type-specific and cell type-shared genes. We found an over-representation of functions linked to signal transducer activity, cell communication, development, behavior, cellular process, enzyme regulator activity, transcription regulator activity and response to stimulus, reflecting processes likely to sculpt cell type-specific profiles. For eQTLs shared in all cell types we found an over-representation of catalytic activity and transport properties (Fisher's exact test p-value<0.05) (Table S5).

We have demonstrated that variants affecting gene regulation act predominantly in a cell type-specific manner, and even cell types closely related as LCLs and T-cells share only a minority of *cis*-eQTLs. We estimate that 69 - 80% of regulatory variants are cell type-specific and that regulatory variant complexity correlates with transcript complexity, implying genotype-specific effects on alternative transcript choice. In addition, cell type-specific eQTLs have smaller effects and tend to localize at greater distances from the TSS, recapitulating enhancer element distributions. The signal of cell type specificity was shown to be primarily due to differential use of regulatory elements of genes that are expressed in almost all cell types. As more tissues are interrogated, we expect diminishing returns in discovery of eQTLs, and it is possible that there is a minimum set of informative tissues for the majority of regulatory variants. Our study highlights the need for extensive interrogation of regulatory variation in multiple cell types and tissues to elucidate their differential functional properties.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge financial support from the Wellcome Trust and NIH to ETD and Infectigen Foundation, Swiss National Science Foundation and, AnEUploidy EU to SEA. Gene expression data is deposite din GEO under accession number GSE17080.

References

1. Dimas AS, et al. PLoS Genet. Oct.2008 4:e1000244. [PubMed: 18974877]
2. Dixon AL, et al. Nat Genet. Oct.2007 39:1202. [PubMed: 17873877]
3. Goring HH, et al. Nat Genet. Oct.2007 39:1208. [PubMed: 17873875]
4. Morley M, et al. Nature. Aug 12.2004 430:743. [PubMed: 15269782]
5. Schadt EE, et al. Nature. Mar 20.2003 422:297. [PubMed: 12646919]
6. Stranger BE, et al. PLoS Genet. Dec.2005 1:e78. [PubMed: 16362079]
7. Stranger BE, et al. Nat Genet. Oct.2007 39:1217. [PubMed: 17873874]
8. Adams MD, et al. Nature. Sep 28.1995 377:3. [PubMed: 7566098]
9. Reymond A, et al. Nature. Dec 5.2002 420:582. [PubMed: 12466854]
10. Su AI, et al. Proc Natl Acad Sci U S A. Apr 2.2002 99:4465. [PubMed: 11904358]
11. Emilsson V, et al. Nature. Mar 27.2008 452:423. [PubMed: 18344981]
12. Myers AJ, et al. Nat Genet. Dec.2007 39:1494. [PubMed: 17982457]
13. Schadt EE, et al. PLoS Biol. May 6.2008 6:e107. [PubMed: 18462017]
14. Campbell DJ, Ziegler SF. Nat Rev Immunol. Apr.2007 7:305. [PubMed: 17380159]
15. Cotsapas CJ, et al. Mamm Genome. Jun.2006 17:490. [PubMed: 16783630]
16. Wu C, et al. PLoS Genet. May.2008 4:e1000070. [PubMed: 18464898]
17. International_HapMap_Consortium. et al. Nature. Oct 18.2007 449:851. [PubMed: 17943122]
18. Stranger BE, et al. Science. Feb 9.2007 315:848. [PubMed: 17289997]
19. www.sciencemag.org - SOM.
20. Heinzen EL, et al. PLoS Biol. Dec 23.2008 6:e1. [PubMed: 19222302]
21. Storey JD, Tibshirani R. Proc Natl Acad Sci U S A. Aug 5.2003 100:9440. [PubMed: 12883005]
22. Goring HH, Terwilliger JD, Blangero J. Am J Hum Genet. Dec.2001 69:1357. [PubMed: 11593451]
23. Ioannidis JP. Epidemiology. Sep.2008 19:640. [PubMed: 18633328]
24. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Nat Genet. Feb.2003 33:177. [PubMed: 12524541]
25. Veyrieras JB, et al. PLoS Genet. Oct.2008 4:e1000214. [PubMed: 18846210]
26. McVean GA, et al. Science. Apr 23.2004 304:581. [PubMed: 15105499]
27. Birney E, et al. Nature. Jun 14.2007 447:799. [PubMed: 17571346]
28. Ashburner M, et al. Nat Genet. May.2000 25:25. [PubMed: 10802651]

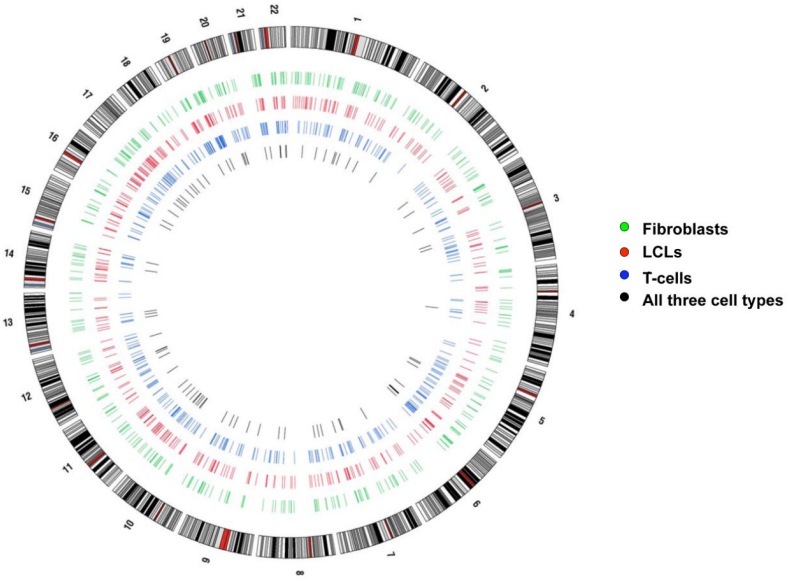

Fig. 1.

Genome-wide map of *cis*-eQTLs in three cell types; *cis*-eQTLs at 0.001 permutation threshold are shown as color-coded lines on their corresponding chromosomal location. Internal black lines represent genes with eQTLs in all cell types.

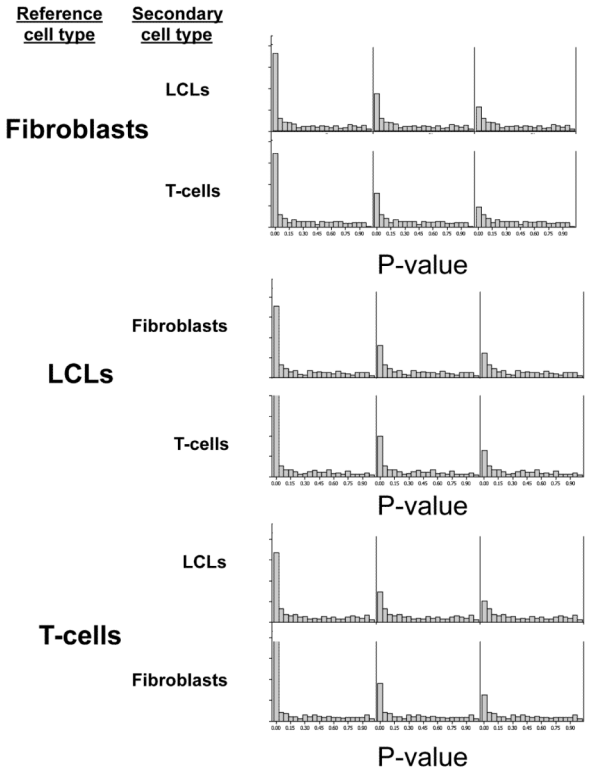

Fig. 2.

SNP-probe pair nominal (uncorrected) p-value distributions for the two secondary cell-types conditional on the reference cell type eQTL, significant at 0.001 permutation threshold are shown. The panels on the horizontal axis correspond (from left to right) to (i) the full p-value distribution of the secondary cell type, (ii) the p-value distribution after excluding significant eQTLs at 0.001 permutation threshold in secondary cell type, and (iii) similar to (ii) at 0.01 permutation threshold.

Europe PMC Funders Author Manuscripts

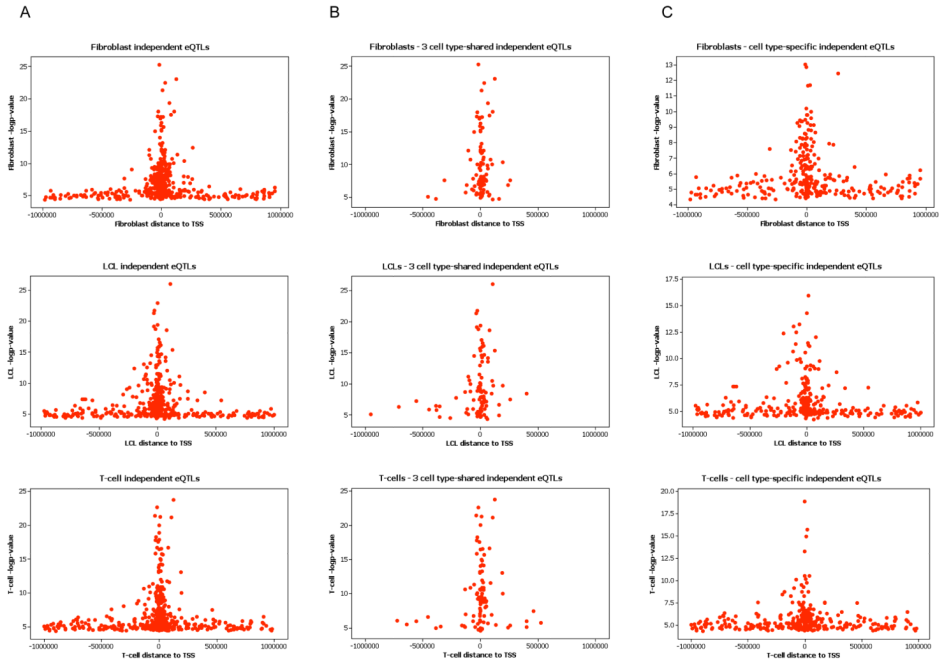

Fig. 3.

Localization of *cis*-eQTLs A) Distance to transcription start site (TSS) of all independent *cis*-eQTLs in each cell type (0.001 permutation threshold). A) shared in all three cell types (0.001 permutation threshold). C) cell type-specific *cis*-eQTLs (0.001 permutation threshold).

Europe PMC Funders Author Manuscripts

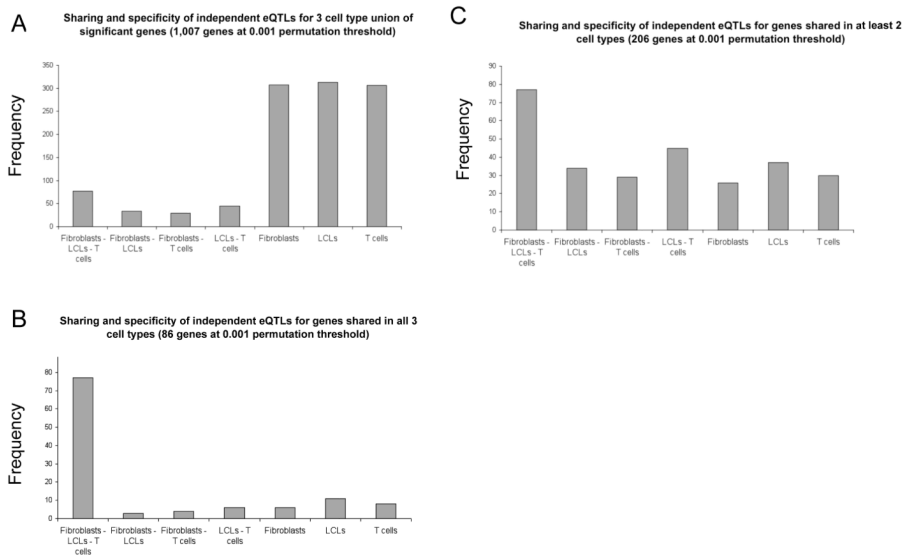

Fig. 4.

Fine-scale overlap of regulatory signals in three cell types A) Cell type-shared and cell type-specific independent *cis*-eQTLs (regulatory intervals) for all genes with a significant association at the 0.001 permutation threshold. B) genes significant in at least two cell types C) genes significant in all three cell types.

Europe PMC Funders Author Manuscripts