Mixed-species RNA-seq for elucidating non-cell-autonomous control of gene transcription

Supplementary Results

Confirming accuracy of FPKM values when performing read separation

To confirm the ease of three-species read separation using the Sargasso pipeline, we took single-species RNA-seq data (rat samples MGLmonoCTR1–3 from ArrayExpress accession E-MTAB-5987 — in this protocol, sample set 6a — human samples GSM2285374–7 from Gene Expression Omnibus series GSE85839 [1] and mouse samples CTR1-34316426, CTR2-34335325, and CTR3-34312414 from ArrayExpress accession E-MTAB-5489 [2]) and calculated the % reads lost for each protein-coding gene expressed at a level >1 FPKM (fragments per kilobase of transcript per million mapped reads, as assessed by normal mapping) when performing the *Sargasso* pipeline requiring disambiguation of reads from the other two species. “Normal mapping” is defined as mapping that requires a perfect match with the target species, but no disambiguation from any other species. For all species’ reads, the overwhelming majority of reads are retained in the overwhelming majority of genes (Supplementary Figure 1a).

For each data set, we also quantified the whole-gene expression level (FPKM) using both a standard protocol (in which reads were first mapped using STAR [3] — requiring a perfect match with the target species, but no disambiguation from any other species — then expression quantified using the Salmon gene and transcript quantification tool [4], and the *Sargasso*-based pipeline (Supplementary Figure 1b–g). The correlation of the results obtained implementing the two approaches is high for all three data sets analyzed (r=0.97–0.99). For example, analysis of a pure rat microglial RNA-seq data set revealed 12,432 genes expressed ≥1 FPKM, using the conventional protocol. Using the *Sargasso* pipeline, genes that lose a lot of reads (due to high rat–mouse or rat–human conservation) will produce a FPKM value that is lower than the actual one. However, this number is small: only 1.6% of the 12,432 genes have an FPKM value that is more than 25% lower than that determined implementing the conventional protocol, and only 0.28% of genes have an FPKM value that is over two-fold lower than that determined implementing the conventional

protocol (Supplementary Figure 1c). We also analyzed the same three data sets quantifying FPKM at transcript level, instead of doing it at whole-gene level. Again, the correlation between the results obtained implementing the two approaches is high for all three data sets analyzed (r=0.95–0.99, Supplementary Figure 2a, c, e). For example, analysis of the pure rat microglial RNA-seq data set revealed 16,169 transcripts expressed ≥1 FPKM, using the conventional protocol. Implementing the *Sargasso* pipeline, only 2.8% of transcripts have an FPKM value whose difference with its counterpart calculated by the conventional protocol is over 25%, and only in 1.1% of transcripts the FPKM value is over two-fold different (Supplementary Figure 2b). Thus, for the human–mouse–rat species combination, implementing the *Sargasso* pipeline does not skew absolute FPKM values substantially in the vast majority of genes or transcripts. Indeed, in samples where gene expression levels can vary by 3–4 orders of magnitude, the relatively modest deviations reported above are not problematic.

Confirming accuracy of fold-change values when performing read separation

In addition to reporting accurate FPKM values, we wanted to confirm that the *Sargasso* approach (applied to a combined human–mouse–rat sample) does not substantially influence the quantification of gene or transcript fold-changes (defined as the fold-difference of a gene's expression in response to a stimulation, relative to a control unstimulated sample). We have taken data generated from a mono-culture of rat microglia, treated or not with inflammation-inducing LPS (E-MTAB-5987-sample sets 6a and 6b from this publication: see below), and mouse (DIV4) and human embryonic stem cell (ESC)-derived neurons treated with high K^+ concentrations to induce membrane depolarization (both E-MTAB-5489)), and calculated *DESeq2* Log_2 fold-change after standard read mapping, and after *Sargasso* read separation (requiring reads to be distinct from the other two species' genomes). The correlation of the results obtained implementing the two protocols is high both at gene level (Rat: r=0.999, Supplementary Figure 3a; Mouse:

r=0.999, Supplementary Figure 3c, Human: r=0.999, Supplementary Figure 3e), and at transcript level (Rat: r=0.997, Supplementary Figure 3b; Mouse: r=0.960, Supplementary Figure 3d, Human: r=0.983, Supplementary Figure 3f). Thus, the *Sargasso* workflow does not substantially affect differential gene expression analysis in human–mouse–rat combined samples.

Testing accuracy of FPKM values when species are evolutionarily close

To assess the accuracy of gene FPKM quantification as species get progressively closer in evolutionary terms, we performed an *in silico* analysis of a data set consisting of 75bp paired-end reads originating from human neural progenitor cells (Gene Expression Omnibus samples GSM2285374–7 [1]), and determined those reads that could be unambiguously attributed to the human genome, when compared to the macaque genome, or the chimpanzee genome (plus mouse and rat for comparison). For the vast majority of genes, the vast majority of human reads can be unambiguously assigned to the human genome, when comparing to the macaque genome (Supplementary Figure 4a). In other words, only a small proportion of human 75bp paired-end reads covered regions of the genome 100% conserved between human and macaque. Moreover, FPKM values calculated after *Sargasso* pipeline implementation for the human genes show an extremely high correlation with FPKM values calculated after conventional mapping, when disambiguating against the macaque data (r=0.987, Supplementary Figure 4d, nearly as high as when separating against rat, r=0.998, Supplementary Figure 4c, and mouse, r=0.999, Supplementary Figure 4b). Even when requiring unambiguous read assignment of human RNA-seq reads when comparing to the chimpanzee genome (median evolutionary separation time of 6.4 million years [5]) more than two-thirds of human genes retain over half their reads when compared with standard mapping (Supplementary Figure 4a), although this loss of ambiguous reads (due to 100% human-chimpanzee sequence conservation) leads to

less accurate FPKM values (r=0.954 compared with standard mapping, Supplementary Figure 4e).

Studying LPS-regulated microglial genes in single species co-cultures

We wanted to confirm that the LPS-induced microglial genes observed to be up- and down-regulated in the presence of neurons and astrocytes were not due to the neurons and astrocytes being from a different species than the microglia but to the presence of LPS. For this purpose, we created an additional set of samples as follows:

Sample	Neurons (rat)	Astrocytes (rat)	Microglia (rat)	LPS
5a	+	+	+	–
5b	+	+	+	+
6a	–	–	+	–
6b	–	–	+	+

We performed differential gene expression analysis between samples 5a and 5b. We took the list of rat microglial genes significantly and greatly (>four-fold) induced (304 genes) or repressed (113 genes) as a consequence of LPS treatment (2a vs. 2b) and then looked at the LPS-dependent regulation of the subset of these genes whose induction could be tracked in a single-species co-culture (5a vs. 5b) by virtue of their expression being >five-fold higher in a pure microglial culture than in the mixed microglia–astrocyte–neuron co-culture (5a vs. 6a), and expressed at least 1 FPKM in mono-cultured microglia. Applying these criteria, we were able, to a first approximation, to monitor the regulation of 108/304 LPS-induced genes, and 44/113 LPS-repressed genes, in microglia in a single-species microglia–neuron–astrocyte co-culture. Reassuringly, these 108 genes were induced (Log2(fold-change): 3.41 ± 0.22, *P*=3.1E-7, Supplementary Figure 5a) by LPS in the single species co-culture, and the 44 genes repressed (Log2(fold-change): –1.43 ± 0.23, *P*=1.2E-4, Supplementary Figure 5b) by LPS in the single species co-culture. Collectively, these data suggest that the expression changes induced by LPS in microglia were not influenced by the presence of cells from different animal species.

References

1 Muffat, J. *et al.* Efficient derivation of microglia-like cells from human pluripotent stem cells. *Nat Med* **22**, 1358-1367, doi:10.1038/nm.4189 (2016).

2 Qiu, J. *et al.* Evidence for evolutionary divergence of activity-dependent gene expression in developing neurons. *Elife* **5**, doi:10.7554/eLife.20337 e20337 [pii] (2016).
3 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:bts635 [pii] 10.1093/bioinformatics/bts635 (2013).
4 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).
5 Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular biology and evolution* **34**, 1812-1819, doi:10.1093/molbev/msx116 (2017).