Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain

Alexandre Kuhn[1,3], Doris Thu[1], Henry J Waldvogel[2], Richard L M Faull[2] & Ruth Luthi-Carter[1]

Human diseases are often accompanied by histological changes that confound interpretation of molecular analyses and identification of disease-related effects. We developed population-specific expression analysis (PSEA), a computational method of analyzing gene expression in samples of varying composition that can improve analyses of quantitative molecular data in many biological contexts. PSEA of brains from individuals with Huntington's disease revealed myelin-related abnormalities that were undetected using standard differential expression analysis.

Many biological processes, including those underlying human diseases, are accompanied by changes of cell populations in corresponding tissues. In the case of neurodegenerative diseases, specific neuronal populations gradually disappear, often paralleled by an increase in glial cell numbers (**Fig. 1a**). Tumor samples can be composed of malignant and nonmalignant cells in varying proportions. The characterization of molecular changes in diseased tissues provides crucial information about pathophysiological mechanisms and is important for the development of targeted drugs and therapies. Identification of these key molecular events, however, can be obscured by tissue heterogeneity and by confounding changes in tissue composition. The goal of the present work was to discern molecular changes in samples of heterogeneous composition without relying on measures of cell content.

We have previously examined gene expression changes in the brains of individuals affected with Huntington's disease[1], an inherited disorder whose pathological hallmark is the degeneration of projection neurons of the caudate nucleus[2]. The predominant disease mechanisms remain unclear. Our previous microarray-based analyses of post-mortem Huntington's disease caudate samples showed extensive gene expression changes in affected individuals compared to controls. However, given neuronal death and glial proliferation in Huntington's disease[3], the detected gene expression changes are bound to also reflect changes in tissue composition. For example, *NEFL* encodes a neuron-specific intermediate filament component whose expression is detected as decreased in brain from individuals with Huntington's disease in a pathological grade–dependent manner (**Fig. 1b**), but it is not possible to distinguish whether this reflects a decreased number of neuronal cells, a decrease of neuron-specific expression or both. In some cases, experimental approaches such as laser-capture microdissection or fluorescence-activated cell sorting can be used to separate these two effects. But these methods have many limitations and cannot be applied in many cases.

We developed a computational method for deconvolving tissue heterogeneity and identifying cell-specific expression changes in the context of cell population shifts (**Fig. 1c**). This method works by exploiting linear regression modeling of queried expression levels to cell type–specific reference measures. A formal statistical description of the method and its detailed implementation is available in Online Methods.

Tissue samples to be analyzed by population-specific expression analysis (PSEA) contain varying proportions of distinct cell populations. In these samples, the contributed expression of a gene in a particular cell type is modeled as being proportional to the size of the cell population. Previous studies have shown that linear models may capture the behavior of the majority of genes[4]. Thus, RNAs expressed in a given population can be detected as expression with a linear dependence on that population's size. Because an accurate measure of population size is often unobtainable (for example, from human clinical or autopsy samples), PSEA instead tracks relative cell population size via the expression levels of marker genes (genes whose expression is highly enriched in a particular cell population in the tissue of interest). Population-specific reference expression signals are constructed by averaging the expression of several marker genes for each population (Online Methods). The selected marker genes should show equivalent cellular expression in the specific conditions to be interrogated. When marker genes are identifiable and appropriately chosen, results derived using PSEA show little dependence on the specific marker choice (Online Methods and **Supplementary Discussion**).

Regression of the expression level of each target gene on population-specific reference measures allowed us to detect individual cell populations expressing that gene (**Fig. 1d**). Regression coefficients yielded normalized population-specific expression levels (Online Methods and **Fig. 1e**). Using RNA mixing experiments, we showed that PSEA can be used to detect expression in one or several populations simultaneously and to quantitatively measure cell population–specific expression (**Supplementary Data 1** and **2**, **Supplementary Figs. 1–8** and **Supplementary Tables 1–3**). Moreover, PSEA enabled statistical testing of differences in population-specific

[1]Laboratory of Functional Neurogenomics, Ecole Polytechnique Fédérale de Lausanne, Switzerland. [2]Department of Anatomy with Radiology and Centre for Brain Research, University of Auckland, New Zealand. [3]Present address: Laboratory of Microfluidics Systems Biology, Institute of Materials Research and Engineering, Singapore. Correspondence should be addressed to A.K. (alexandre.m.kuhn@gmail.com).

RECEIVED 23 MAY; ACCEPTED 4 AUGUST; PUBLISHED ONLINE 9 OCTOBER 2011; DOI:10.1038/NMETH.1710

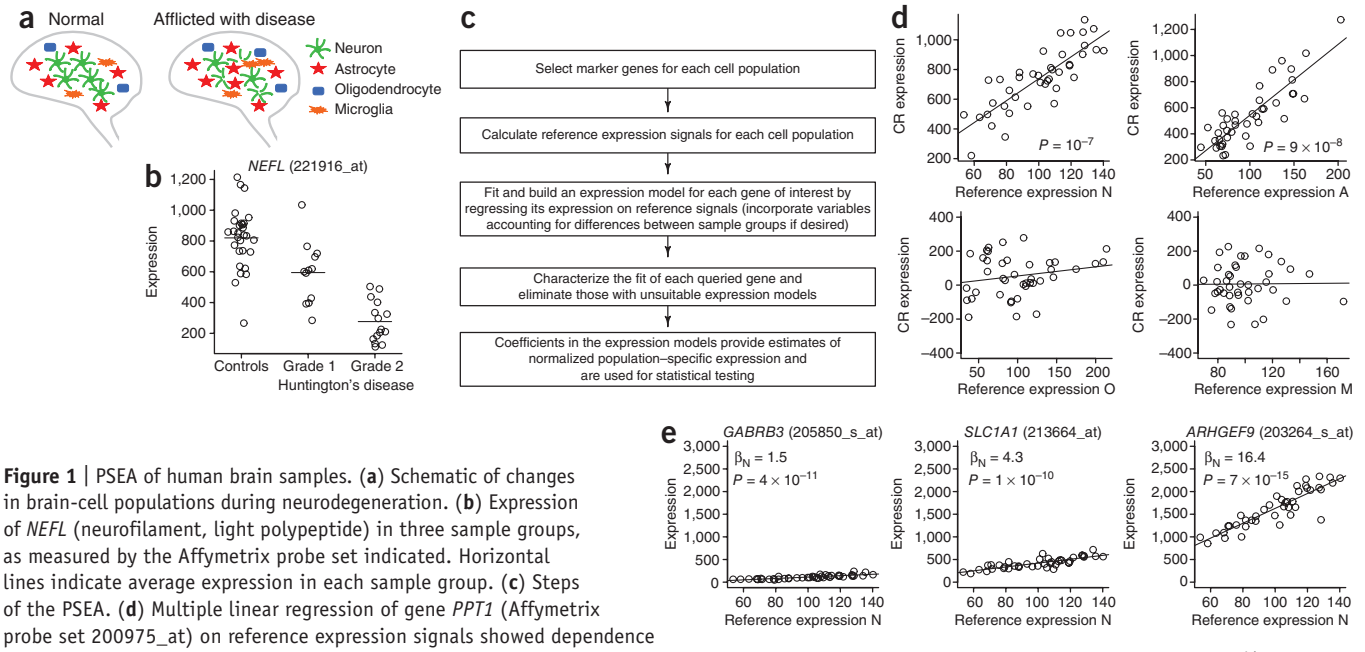

Figure 1 | PSEA of human brain samples. (**a**) Schematic of changes in brain-cell populations during neurodegeneration. (**b**) Expression of *NEFL* (neurofilament, light polypeptide) in three sample groups, as measured by the Affymetrix probe set indicated. Horizontal lines indicate average expression in each sample group. (**c**) Steps of the PSEA. (**d**) Multiple linear regression of gene *PPT1* (Affymetrix probe set 200975_at) on reference expression signals showed dependence on neuronal (N) and astrocytic (A) reference signals but not on oligodendrocytic (O) and microglial (M) signals. CR, component plus residual[14]. Population-specific expressions with $P < 0.05$ are indicated. (**e**) Linear regression of the expression of three genes on the neuronal reference signal. β_N, slope of the regression and relative neuron-specific expression.

expression across sample groups (for example, disease versus control; Online Methods). Thus, the change in expression of an RNA in a given cell type could be determined, despite a systematic difference in that cell type's relative abundance.

In keeping with our original goal of using PSEA to better understand molecular events in neurodegenerative disease brains, we applied PSEA to gene expression profiles obtained from caudate nucleus samples from individuals with Huntington's disease, and age- and gender-matched controls[1] (**Supplementary Figs. 9–13**, **Supplementary Tables 4–10** and **Supplementary Data 3–5**). In this brain region, Huntington's disease is characterized by neuronal loss and parallel increases of astrocytes and microglia. Upon preliminary analysis of control samples alone, we found that individual variability of the neuronal, astrocytic, oligodendrocytic or microglial populations was enough to account for the variability of the vast majority of probe sets (Online Methods). We thus used signals from these four cell types to construct expression models and estimate population-specific expression. PSEA assigned 1,802 probe sets an expression model with the required goodness of fit (Online Methods). Among these, we detected various patterns of specific gene expression in the four brain-cell populations (**Supplementary Fig. 14**) over a wide range of population-specific expression levels (**Fig. 1e**).

We also detected differences in population-specific expression between control and Huntington's disease samples, observable as different slopes of the regression lines in the two sample groups (**Fig. 2**). For example, neuron-specific expression of *PPP3CA* (**Fig. 2c**) was smaller for the Huntington's disease sample (grade 1) as compared to control samples. This is in contrast to *WASF1* (**Fig. 2a**), whose neuron-specific expression did not change between the two sample groups, and *ATP6V1A* (**Fig. 2e**), which showed a neuron-specific increase in the Huntington's disease sample. We also detected different specific expression in oligodendrocytes (**Fig. 2i,k**) and in other brain-cell populations (**Supplementary Fig. 15**, **Supplementary Tables 11–15** and Online Methods).

We compared population-specific expression detected with PSEA to the results of a standard differential analysis of (total) expression in the same sample set. For instance, total expression of *WASF1* was lower in Huntington's disease samples (**Fig. 2b**) owing to their smaller neuronal content. In this case, the decreased number of neurons in Huntington's disease tissue samples sufficed to explain the observed change in total expression (**Fig. 2a**). In the case of *PPP3CA*, a large decrease in total expression (**Fig. 2d**) was deconvolved by PSEA into two effects: a decrease in the number of neuronal cells in the Huntington's disease samples together with a Huntington's disease–related decrease in neuronal gene expression. In a notable example, total expression of *ATP6V1A* was decreased in Huntington's disease samples (**Fig. 2f**), whereas PSEA detected increased neuron-specific expression (**Fig. 2e**). Thus, standard analyses may return misleading findings when the magnitude of the change in the cell population is greater than the change in specific expression. These examples illustrate the ability of PSEA to differentiate between a change in neuronal cell–specific expression from a change owing to population effects. We systematically compared PSEA and standard analysis (**Supplementary Data 6**, **Supplementary Figs. 16–18** and **Supplementary Table 16**).

In addition to correcting for systematic changes in cell populations, PSEA was also effective in accounting for sample-to-sample variations. The large variability of the oligodendrocytic fraction (**Supplementary Fig. 9**) resulted in increased (total) expression variability and effectively prevented the detection of expression changes by standard analysis (**Fig. 2g–l**). Using PSEA, we detected 205 genes that had different oligodendrocyte-specific expression in Huntington's disease samples versus controls (**Supplementary Tables 11b** and **13**). Many of the genes differentially expressed in oligodendrocytes were associated with myelin formation and stability, suggesting important modifications in myelination pathways in brains of individuals with Huntington's disease. We discuss expression changes detected with PSEA in the various brain-cell populations in **Supplementary Data 4**.

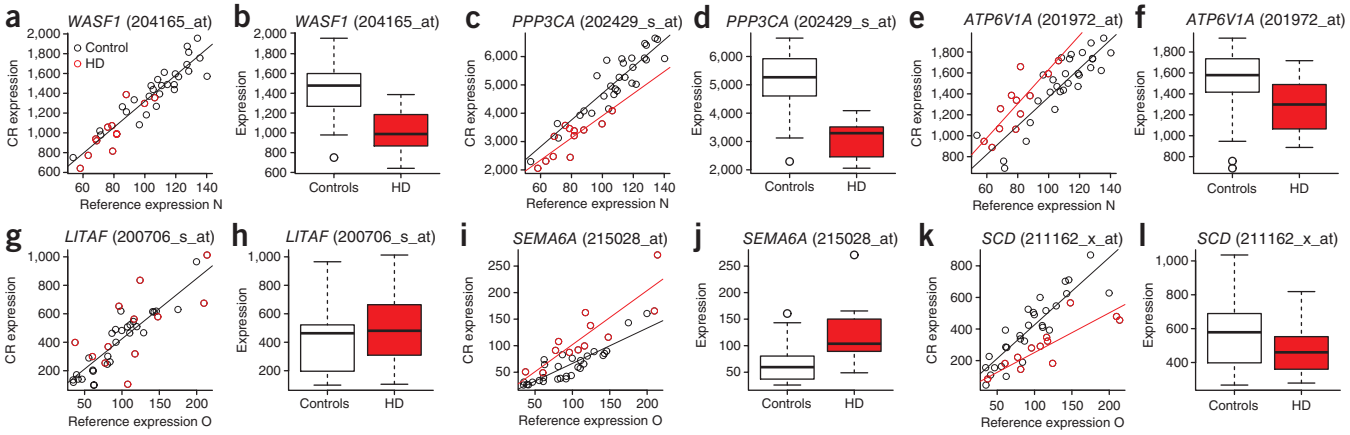

Figure 2 | Comparison of population-specific and total expression change in tissue samples with cell-composition changes. (**a**,**b**) Neuron-specific (**a**, modeled as identical in the two sample groups) and total (**b**, log fold change –0.5) expression of *WASF1*. (**c**,**d**) Neuron-specific (**c**, log fold change –0.3) and total (**d**, log fold change –0.8) expression of *PPP3CA*. (**e**,**f**) Neuron-specific (**e**, log fold change 0.2) and total (**f**, log fold change –0.2) expression of *ATP6V1A*. (**g**,**h**) Oligodendrocyte-specific (**g**, modeled as identical in the two sample groups) and total (**h**) expression of *LITAF*. (**i**,**j**) Oligodendrocyte-specific (**i**, log fold change 0.6) and total (**j**, log fold change 0.8) expression of *SEMA6A*. (**k**,**l**) Oligodendrocyte-specific (**k**, log fold change –0.8) and total (**l**, log fold change –0.2) expression of *SCD*. Huntington's disease (HD) samples and controls were compared. Gene expression was measured by the Affymetrix probe sets indicated in parentheses. Reference expression N and O are the neuronal and oligodendrocytic reference expression signals, respectively.

To validate expression changes detected by PSEA, we assessed the expression of several of the corresponding proteins by immunohistochemistry and confirmed PSEA-based predictions (**Supplementary Data 5**, **Supplementary Figs. 15**, **19** and **20** and **Supplementary Table 17**). Finally, we tested the applicability of PSEA in a tissue other than brain by deconvolving whole-blood expression profiles obtained from kidney-transplant recipients[5] (**Supplementary Data 7**). We validated PSEA-derived population-specific expression using independent expression profiles from isolated blood-cell populations (**Supplementary Fig. 21**) and detected transplant rejection–related expression changes in granulocytes (**Supplementary Table 18**).

A handful of previous studies have addressed the problem of sample heterogeneity in gene-expression analyses[4–9], a major outstanding issue for the analysis of transcriptomic and other types of high-dimensional data. Three previous studies have treated the estimation of differential expression in the presence of varying sample composition, relying either on information on sample composition[5,10] or known expression profiles of component cells[11]. The major advantage of our strategy is its ability to estimate population-specific expression without these types of external information. The benefits of such a method include its wide applicability, including retrospective implementation and reliance on data from one single assay (**Supplementary Discussion**).

We identified a large set of myelin-related expression changes, suggesting modified myelin synthesis and composition in oligodendrocytes of individuals affected with Huntington's disease. These results are particularly intriguing when considered together with recent imaging studies describing early white matter abnormalities in the brain of individuals with Huntington's disease[12] consistent with myelin breakdown[13]. Although it is not yet possible to assign causality, these results allowed us to propose new molecular hypotheses regarding the involvement of oligodendrocytes and myelin in the etiology of Huntington's disease.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. Gene Expression Omnibus: GSE19380.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was funded by the Huntington's Disease Society of America, the Novartis Foundation and the Swiss National Science; Foundation. We thank L. Glauser for technical assistance; F. Brunet, S. Lengacher, I. Allamand, M. Marti, L. Claivaz, M. Delorenzi, T. Sengstag, D. Goldstein and S. Brahmachari for helpful discussions; B. Deplancke, J. Rougemont, I. Krier and M. Nalls for critically reading the manuscript; members of Vital-IT and the Swiss Institute of Bioinformatics for providing computing infrastructure, and members of the Lausanne DNA Array facility for assistance in processing microarray samples from primary cell cultures.

AUTHOR CONTRIBUTIONS

A.K. developed and applied PSEA, performed experiments with cultured cells, and wrote the manuscript, with input from R.L.-C; D.T., H.J.W. and R.L.M.F. performed immunohistochemical experiments with brain sections; and R.L.-C. conceptualized the project and directed the study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Hodges, A. *et al. Hum. Mol. Genet.* **15**, 965–977 (2006).
2. Reiner, A. *et al. Proc. Natl. Acad. Sci. USA* **85**, 5733–5737 (1988).
3. Vonsattel, J.P. *et al. J. Neuropathol. Exp. Neurol.* **44**, 559–577 (1985).
4. Lu, P., Nakorchevskiy, A. & Marcotte, E.M. *Proc. Natl. Acad. Sci. USA* **100**, 10370–10375 (2003).
5. Shen-Orr, S.S. *et al. Nat. Methods* **7**, 287–289 (2010).
6. Venet, D., Pecasse, F., Maenhaut, C. & Bersini, H. *Bioinformatics* **17** (suppl. 1) S279–S287 (2001).
7. Stuart, R.O. *et al. Proc. Natl. Acad. Sci. USA* **101**, 615–620 (2004).
8. Lähdesmäki, H., Shmulevich, L., Dunmire, V., Yli-Harja, O. & Zhang, W. *BMC Bioinformatics* **6**, 54 (2005).
9. Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H.F. *PLoS ONE* **4**, e6098 (2009).
10. Ghosh, D. *Bioinformatics* **20**, 1663–1669 (2004).
11. Wang, M., Master, S.R. & Chodosh, L.A. *BMC Bioinformatics* **7**, 328 (2006).
12. Ciarmiello, A. *et al. J. Nucl. Med.* **47**, 215–222 (2006).
13. Rosas, H.D. *et al. Neuroimage* **49**, 2995–3004 (2010).
14. Fox, J. *Applied Regression Analysis, Linear Models, and Related Methods* (Sage Publications, Inc., 1997).

ONLINE METHODS

Estimating population-specific expression from composite samples. Let us consider gene expression measured from a sample composed of P cell populations. We assume that the measured expression y is the sum of each population's expression and a constant background term. A single population's contribution is proportional to its size and population-specific expression. The measured expression thus is

$$y = a + \sum_{p=1}^{P} x_p f_p, \tag{1}$$

where a is the background, x_p the specific expression of the gene of interest in population p and f_p the fraction of population p in the composite sample. Specific expression x_p thus is the (background-subtracted) measure of gene expression obtained from a sample composed of population p only.

Consider a set of I composite samples of varying composition of known populations. Equation (1) suggests that we can estimate specific expressions x_p by regressing measured expression y on f_p according to the following linear statistical model

$$y_i = a + \sum_{p=1}^{P} x_p f_{p,i} + \varepsilon_i, \tag{2}$$

where ε_i is the standard error term of linear models and $i = 1, \ldots, I$. x_p is uniquely determined if $I > P$. When sample composition is unknown, we propose to use the expression of marker genes as surrogate of population size. We define a marker gene for population p^* as a gene expressed in population p^* only. Its specific expression thus is

$$x_p = \begin{cases} x_{p^*} > 0, & p = p^* \\ 0, & p \neq p^* \end{cases} \tag{3}$$

and the expression (see equation (1)) of a marker gene for population p^* in a composite sample i simplifies to $y_{p^*} = a + x_{p^*} f_p$. Solving for f_p and replacing the corresponding expression in equation (1) we obtain the following (nonlinear) expression model

$$y = a\left(1 - \sum_{p,p^*=1}^{P} \beta_p\right) + \sum_{p,p^*=1}^{P} \beta_p y_{p^*} \tag{4}$$

where $\beta_p = x_p / x_{p^*}$, the specific expression of the gene of interest in population p normalized to the specific expression of the population marker. If we make the additional assumption that measured marker expression can be modeled without background term ($y_{p^*} \approx x_{p^*} f_p$, that is, assuming that marker expression is high compared to the background), the expression model (equation (1)) becomes

$$y = a + \sum_{p,p^*=1}^{P} \beta_p y_{p^*} \tag{5}$$

indicating that we can obtain approximate estimates of relative specific expression β_p by linearly regressing measured expression y on the expression of marker genes y_{p^*}. To reduce noise in marker gene expression measures y_{p^*}, we used averaged expression of several marker genes for each population. Normalized population-specific expression becomes $\beta_p = x_p / E(x_{p^*})$, the specific expression of the gene of interest in population p normalized to the average specific expression of population P's selected marker genes. In the presentation of Results, we alternatively call the averaged expression of several marker genes (for a given population) the 'population reference signal'.

We now consider two sample groups (for example, control and disease) with possibly different specific expression levels (denoted x_p for group 1 and x'_p for group 2). We are interested in estimating the change of specific expression between groups ($x'_p - x_p$). We use an indicator variable d_i

$$d_i = \begin{cases} 0, & \text{if sample } i \text{ is in group 1} \\ 1, & \text{if sample } i \text{ is in group 2} \end{cases} \tag{6}$$

and model gene expression in composite samples as

$$y_i = a + \sum_{p,p^*=1}^{P} \beta_p y_{p^*,i} + \sum_{p,p^*=1}^{P} \beta'_p (y_{p^*,i}\, d_i) \tag{7}$$

where $\beta'_p = (x'_p - x_p)/x_{p^*}$ is the relative difference between specific gene expression in the two groups and expression $y_{p^*,i}\, d_i$ represents an interaction regressor constructed by multiplying $y_{p^*,i}$ and d_i. For samples of group 1, the model thus reduces to equation (5) and for samples of group 2, it becomes

$$y_i = a + \sum_{p,p^*=1}^{P} (\beta_p + \beta'_p) y_{p^*,i} \tag{8}$$

Note that $\beta_p + \beta'_p = x'_p / x_{p^*}$, the relative specific expression in group 2. Simultaneous regression of measured expression y_i on the expression of marker genes y_{p^*} in the two sample groups (equation (7)) allows us to test the null hypothesis H_0: $\beta'_p = 0$, which is equivalent (because of scale invariance) to testing H_0: $x'_p - x_p = 0$, the null hypothesis of equal population p specific expressions in the two sample groups. The estimate of the population-specific fold change in expression in group 2 is $(\beta_p + \beta'_p)/\beta_p$.

When applying this method to our Huntington's disease dataset, we also considered both Huntington's disease grade 1 and 2 cases simultaneously and aimed at estimating population-specific changes in these 2 groups compared to control cases. We thus had 3 sample groups and used the expression model

$$y_i = a + \sum_{p,p^*=1}^{P} \beta_p y_{p^*,i} + \sum_{p,p^*=1}^{P} \beta'_p (y_{p^*,i}\, d_{1,i}) + \sum_{p,p^*=1}^{P} \beta''_p (y_{p^*,i}\, d_{2,i}) \tag{9}$$

where $d_{1,i}$ and $d_{2,i}$ are indicator variables for grade 1 and grade 2 samples, respectively, and $\beta'_p = (x'_p - x_p)/x_{p^*}$ and $\beta''_p = (x''_p - x_p)/x_{p^*}$ are the relative differences between specific gene expression in the grade 1 and the controls, and between grade 2 and the controls, respectively.

Gene expression datasets. We first aimed at validating PSEA-generated estimates of population-specific expression levels (β_p in equation (5)) using a set of samples composed of cell populations with known specific expressions. We cultured rat primary neurons, astrocytes, oligodendrocytes and microglia separately and extracted their RNA to generate expression profiles for these

doi:10.1038/nmeth.1710

four cell types using Affymetrix Rat 230 2.0 microarrays. In addition, we mixed these reference RNAs to generate a collection of composite samples that we submitted to gene expression profiling as well. This gene expression dataset was comprised of 26 profiles: 4 replicates for each of the 4 cell types and 10 mixed RNA samples (**Supplementary Table 1** lists samples and mixing proportions). We performed PSEA on 8 composite samples (samples 17–24). Given the small number of mixed samples, we excluded the additional samples comprising microglial RNA (samples 25 and 26) to restrict the number of coefficients to be estimated.

In a second application, we reanalyzed data from our previous gene expression study of human Huntington's disease brain[1]. In this study, we profiled 33 control samples, 14 Huntington's disease grade 1 samples and 18 grade 2 samples on Affymetrix HG-U133 A and B microarrays. We dropped 2 control, 1 Huntington's disease grade 1 and 3 Huntington's disease grade 2 samples upon microarray quality assessment[1]. Additionally, 2 controls and 1 Huntington's disease grade 1 samples were deemed statistical outliers and we thus applied PSEA on the remaining 29 controls, 12 Huntington's disease grade 1 and 15 Huntington's disease grade 2 samples (listed in **Supplementary Table 4**). RMA was used to normalize microarrays and perform gene expression summarization[15].

Marker genes and population reference signals. For each cell population known (artificially mixed RNA dataset) or suspected (Huntington's disease dataset) to contribute to gene expression, we interrogated known marker genes. We avoided human genes known to be dysregulated in Huntington's disease, since such changes would systematically bias PSEA. Many marker genes are probed by several probe sets on Affymetrix microarrays and we selected probe sets as follows for creating population reference signals: we eliminated nonselective probe sets by verifying their transcript mapping using Adapt[16]. We also eliminated probe sets with a correlation coefficient to other probe sets (annotated with the same gene) lower than 0.7 as probe sets measuring the same signal should show strongly correlated signals (and obvious non-correlation might comprise dysregulated marker genes). Marker genes and probe sets selected for various cell populations are shown in **Supplementary Table 2** for mixed RNA samples and in **Supplementary Table 5** for human Huntington's disease dataset. Reference signals for each population were constructed as follows: First, each probe set was given an equal weight by normalizing it to an average value of 100. Probe sets reporting the same marker gene expression were then averaged to obtain marker gene expression measures. Finally, we averaged marker gene expression measures within each cell population to obtain population-specific reference signals. These were used as independent variables in regression (y_{p^*}, in equations (5), (7) and (9)).

Where candidate cell type–specific markers need to be defined *de novo*, several approaches can be taken. Where database or literature-based information is not available, tissue-based approaches using and potentially combining histochemical stains, immunohistochemistry, microdissection and/or gene expression profiling can be used to identify cell type–specific markers. In some instances, tissues may be dissociated and pools of individual cells may be selected by molecular, morphologic or functional criteria and subsequently be subjected to gene expression profiling. This may be greatly facilitated by using labeled cells from reporter gene transgenic mice or cells immunolabeled for cell surface markers.

Variable subsets for gene-expression modeling. Given the small sample size of our datasets, we aimed to limit the number of contributing populations included in the expression model of each probe set (that is, the number of predictor variables in regression, to avoid overfitting). We thus implemented a variable selection procedure based on least-squares fitting of all allowed variable subsets. In short, we determined a maximal number of regressors given our sample size and fitted all possible variable subsets. In the case of the detection of specific expression differences in Huntington's disease versus control, we assumed that one population at most showed a change. Alternative, less exhaustive but faster variable selection procedures can also be used, for example, stepwise methods (see our application of PSEA to the deconvolution of expression in blood in **Supplementary Data 7**). Predictor variables used in statistical model building on the mixed RNA dataset corresponded to the cell populations used for generating composite samples. Since we applied PSEA on mixed samples composed of RNAs from neuronal, astrocytic and oligodendrocytic cultures, we used the corresponding three population reference signals as independent variables in PSEA. The set of fitted models is shown in **Supplementary Table 3**. For PSEA on the Huntington's disease dataset, we chose not to assume a priori knowledge of populations contributing to (total) expression. To determine the set of reference signals (and thereby contributing populations) needed to model expression for the majority of probe sets, we regressed probe set signals from control samples with reference signals representing known brain cell populations (thereby covering most probable sources of gene expression in homogenate brain samples; **Supplementary Table 5**). Given the modest number of control samples, we constrained variable subsets to a maximum of three reference signals, resulting in a total of 130 allowed variable subsets (**Supplementary Table 6**). Upon variable selection a minority of probe sets was assigned an expression model with one or more predictor variable accounting for expression in erythrocytes (and their precursors, that is, reticulocytes), endothelial cells, fibroblasts, pericytes or smooth muscle cells (compared to probe sets fitted with variables for neurons, astrocytes, oligodendrocytes or microglia, **Supplementary Table 10**). We thus dropped these predictor variables in further analyses. We used reference signals corresponding to neuron, astrocyte, oligodendrocyte and microglia populations to fit expression in control and Huntington's disease samples simultaneously and detect genes showing differential population-specific expression across sample groups. Regressors accounting for changes in population-specific expression ($y_{p^*} d$ in equations (7) and (9)) are marginal to their corresponding population-specific marker (y_{p^*}) and variable subsets containing $y_{p^*} d$ without y_{p^*} were omitted. Further, regressors capturing population-specific differences in expression ($y_{p^*} d$ in equations (7) and (9)), were moderately correlated. To avoid additional collinearity among regressors and the associated detrimental effects on statistical fit, we restricted variable subsets to those with one such regressor ($y_{p^*} d_1$) when fitting control and Huntington's disease grade 1 samples. When testing for changes in Huntington's disease grade 1 and grade 2 samples (fitting control, Huntington's disease grade 1 and grade 2 samples), we restricted to variable subsets with at most two regressors coding for specific expression changes in the same population (for example, $y_{p^*} d_1$ and $y_{p^*} d_2$). Again, because of the limited number of samples, we limited variable subsets to

those with a maximum of four regressors. To summarize, using four possible contributing cell populations (neurons, astrocytes, oligodendrocytes, microglia) and allowing models with no more than four regressors (excluding the intercept term), and additionally assuming that specific expression changes (between control and Huntington's disease groups) occur in one cell population at most, we limited the set of statistical models to 44 allowed models to account for expression in control and Huntington's disease grade 1 samples (**Supplementary Table 7**), and to 88 allowed models to account for expression in control, Huntington's disease grade 1 and grade 2 samples (**Supplementary Table 8**).

Model selection procedure. For each probe set, we performed ordinary least-square fitting of all allowed variable subsets and selected the best model using Akaike's AIC ('an information criterion')[17]. Instead of selecting the model with smallest AIC, we considered all models with AIC measures close to the smallest AIC because it allowed us to assess the specificity of the selected model or detected expression change. Models within two AIC units of the model with lowest AIC were considered to describe expression data equally well[18]. Among those, we selected the model with the lowest number of regressors (and with the lowest AIC in case several models within 2 AIC units shared the same number of regressors). All selected models (1 for each probe set) were further characterized and those that did not qualify as expression models were filtered out. Specifically, we eliminated models with significantly negative specific expression coefficients ($P < 0.05$). In addition, models containing a regressor accounting for an expression change (that is, either $y_{p^*} d_1$ or $y_{p^*} d_2$ or both) were tested for non-negative specific expression in Huntington's disease sample groups: Given our parametrization (equations (7) and (9)), we thus tested the linear hypotheses H_0: $\beta_p + \beta'_p = 0$ and/or H_0: $\beta_p + \beta''_p = 0$ (that is, specific expression in Huntington's disease grade 1 different from 0 and/or specific expression in Huntington's disease grade 2 different from 0) using an F-test ($P < 0.1$). We interpreted models with large fitted intercepts as evidence of an expression source not represented by any of our chosen reference signals (or, alternatively, the consequence of a large "error-in-variable" effect) and these were also eliminated from further consideration. The expression model in equation (4) suggests that the part of expression that is not dependent on population markers, that is

$$a\left(1 - \sum_{p,p^*=1}^{P} \beta_p\right)$$

and which corresponds to the intercept term in our regression, has an upper bound given by the background of expression measure and is inversely related to the number of populations contributing to expression. In practice, however, we observed that fitted intercept terms were moderately correlated with mean probe set expression (data not shown). Therefore, to avoid predominantly filtering out probe sets with large mean expression (resulting from the use of a fixed threshold on fitted intercepts), we used a relative criterion and eliminated probe sets with a ratio of fitted intercept over mean expression greater than 0.5. Finally, we discarded probe sets whose response variability was poorly explained by the selected expression model and filtered out selected models with an adjusted R^2 lower or equal to 0.6. Plots of adjusted R^2 versus intercept/mean were useful to perform a general assessment of the statistical models obtained (**Supplementary Fig. 12**) and helped us define appropriate levels for these two selection criteria. **Supplementary Table 9** summarizes the fit quality criteria described above and shows how many probe sets passed each.

Characterization of differential population-specific expression. We considered probe sets fitted with a regressor accounting for a change in specific expression as showing evidence of differential population-specific expression. Estimates of population-specific fold changes in expression between control and grade 1 or grade 2 samples are given by $(\beta_p + \beta'_p)/\beta_p$ and $(\beta_p + \beta''_p)/\beta_p$, respectively. In addition to the statistical significance of the change (given by the p-value associated with coefficients β'_p or β''_p, see equations (7) and (9)), we aimed to estimate the confidence that the change occurred in a particular population (and not in any other population). This is particularly relevant since we constrained expression models to those accounting for a single change in population-specific expression only. Moreover, regressors accounting for specific expression changes showed moderate correlation and a confidence factor could help to characterize the population specificity of the detected change. For each selected model, we defined a confidence factor as the number of models accounting for a change in the same population as the selected model (for PSEA on Huntington's disease grade 1 and grade 2 simultaneously, the change could be detected in the same or in the other Huntington's disease sample group) among all 'equivalent' models (that is, those within 2 AIC units of the model with lowest AIC) over the number of 'equivalent' models. The confidence factor thus is 1 if all of the 'equivalent' models contain the same regressor as the selected model. It is close to 0 if there are many 'equivalent' models and only a small fraction account for a change in the same population as the selected model. Confidence factors around 1 thus supported that the detected change was specific for a particular population whereas the reverse was true for confidence factors around 0. Also, it is worth considering the number of 'equivalent' models for any selected model, since this indicates how good a particular model is at describing the data compared to any other allowed model and thus is informative of model specificity.

Outlier detection. We used Cook's distance (Di[19]) to detect individual samples having a strong influence on data fits. Specifically, we calculated Di for all probe sets with expression models that passed our fit quality criteria. Samples with systematically large average Di were flagged. This was repeated for all applications of PSEA. Samples H111 and HC68 were found to have outstanding mean Di and were thus discarded. Plots of marker gene expression across samples were also used to spot unusual samples. Sample 18 displayed an unusually large value for microglial marker expression and showed large Di for a moderate fraction of selected probe sets; it was also discarded from the analysis.

Fit diagnostic and characterization. Compliance with three assumptions of linear least-squares fitting was investigated: normally distributed errors, constant error variance and linearity. Error distributions were deemed to be normal by comparing the sample distribution of studentized residuals with quantiles

doi:10.1038/nmeth.1710

of the normal distribution (quantile-quantile plot), for a large number of fitted probe sets. Similarly, we examined studentized residuals versus fitted responses for a large number of expression models. Most probe sets had constant error variance. A minority showed increasing error variance with increasing fitted expression values. This increase was modest and was deemed not to compromise ordinary least-squares-based coefficient estimation. Finally, model linearity was checked by looking at partial residual plots[14]. The vast majority of fitted probe sets did not reveal clear nonlinearity.

Differential analysis of total expression. We performed differential expression analysis of Huntington's disease grade 1 versus control samples using a *t*-test on $\log_2$-transformed expression values. Adjusted p-values were obtained using Benjamini-Hochberg correction[20].

Computer implementation. PSEA and differential expression analysis were implemented in R (R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2009; http://www.r-project.org/). We used the Bioconductor package 'affy' for microarray data normalization and gene expression summarization[21]. Statistical analyses were implemented using the standard R function 'lm' (for both PSEA and standard differential analysis) and 'extractAIC' was used to obtain AIC values. We used the function 'linear.hypothesis' in the 'car' package (Fox, J. car: Companion to Applied Regression, version 1.2-7, 2007; http://socserv.socsci.mcmaster.ca/jfox/) to perform linear hypothesis tests. An alternative solution to implement PSEA in R is to perform a stepwise model selection (**Supplementary Data 7**).

Cell cultures and RNA extraction. Primary cultures were prepared in accordance with the European Community directive for the care and use of laboratory animals (86/- 609/-EEC) and the Swiss Academy of Medical Science and were authorized by the veterinary office of the canton of Vaud (authorization 1667.2). Rat primary neuronal and glial cells were isolated from cerebral cortices of postnatal day-1 pups. After dissection of the cortices, enzymatically dissociated cells[22] were plated in B27 neurobasal medium (NB) or supplemented Eagle's nasal medium (BME-C) to obtain neuronal or mixed glial cultures, respectively. The latter was used to isolate microglia, oligodendrocytes and astrocytes based on a modified version of a shaking protocol[23]. In short, upon confluence the mixed culture was shaken for 1 h on an orbital shaker at 200–250 r.p.m. to detach microglial cells. The supernatant was then collected to start a microglial subculture and the mixed culture was shaken on overnight after replacement of the culture medium. This prolonged shaking resulted in an enrichment of oligodendrocyte progenitor cells in the supernatant and this supernatant was used to start an oligodendrocyte subculture by plating detached cells in Dulbecco's Modified Eagle's Medium-Ham's F12 (DMEM-F12). The cells that remained attached at the bottom of the flask were predominantly astrocytes, and they were replated in a new dish and cultured with BME-C. Total RNA was extracted from the different subculture using RNeasy Mini kit (Qiagen) following the manufacturer's protocol. Neuronal RNA was extracted 2 weeks after the initial seeding. Microglial, oligodendrocytic and astrocytic RNA was extracted 2, 1 to 5 and 3 d after the specific subculture was initiated, respectively.

Immunocytochemistry and characterization of primary neural cultures. Standard cell type markers were used to characterize neuronal and glial cultures. Isolectin B4 (Sigma-Aldrich) and antibodies to NeuN (Chemicon), O4 (Boehringer Mannheim) and GFAP (Dako) were used to identify microglia, neurons, oligodendrocytes and astrocytes, respectively. Photomicrographs of a neuronal culture, a mixed glial culture, microglia, oligodendrocytes and astrocytes under visible light and after fluorescent immunolabeling are shown in **Supplementary Figure 1**. The presence of different cell types in each cell culture was detected by double labeling. An example is shown in **Supplementary Figure 2**; here we used antibodies to GFAP in addition to antibodies to each specific cell marker to detect the presence of astrocytes among neurons, microglia or oligodendrocytes. Alternatively, we used 4′,6-diamidino-2-phenylindole (DAPI) (nonspecific nuclear DNA staining) to differentially reveal unlabeled cells. As an example, in **Supplementary Figure 2** we show the presence of a cell that did not express GFAP among GFAP-expressing astrocytes. Combining these two approaches, we could precisely assess the cell-specific enrichment we achieved by counting the fraction of cells of interest in each culture before proceeding to RNA extraction. At least 80–90% enrichment was reached for all cultures except for oligodendrocytes, which showed 70–80% enrichment.

Immunohistochemistry in human brain sections. Brain tissue from eight individuals with Huntington's disease and eight individuals with no history of neurological disease (controls) was obtained from the Neurological Foundation of New Zealand Human Brain Bank, Centre for Brain Research, University of Auckland, and informed consent was obtained from all families (**Supplementary Table 17**). Fifty-micrometer perfused-fixed, coronal striatal human sections were processed free-floating in tissue culture wells. The sections were washed in PBS and 0.2% (vol/vol) Triton X (PBS-triton), incubated for 20 min in 50% methanol and 1% H_2O_2, washed (three times for 10 min) in PBS-triton and then incubated with primary antibodies for 2–3 d on a shaker at 4 °C. The following antibodies were used: rabbit antibody to PPM1H (Strategic Diagnostics; 1:1,000), mouse antibody to PGM2L1 (Abnova; 1:1,000), mouse antibody to ID4 (Millipore; 1:500), mouse antibody to HIBADH (Abnova; 1:500), rabbit antibody to GFAP (Dako; 1:5,000). The sections were then washed (three times for 10 min, PBS-triton) and incubated overnight with a biotinylated rabbit or mouse immunoglobulin (H+L) secondary antibody (diluted 1:200). After washing (three times for 10 min, PBS-triton), the sections were incubated for 1 h at room temperature (21 °C) in avidin-biotin complex (ABC) mix and exposed to 0.5% (vol/vol) 3,3-diaminobenzidine tetrahydrochloride (DAB) solution (metal-enhanced DAB substrate kit, Thermo Scientific) for 10 min to produce a brown reaction product. The sections were then washed in PBS-triton, mounted on gelatin-coated slides, rinsed in distilled water, dehydrated through a graded alcohol series to xylene, and then a drop of mounting medium (DPX) was applied onto the coverslip and the coverslip was placed carefully onto the dehydrated section.

Immunohistochemical image analysis. To analyze the immunostaining intensity of *PPM1H*- and *PGM2L1*-immunostained cells in caudate nucleus sections, eight non-overlapping images were taken at 20× from the caudate of each control and Huntington's disease sample. Using the image analysis program ImageJ, boxes were drawn within the cell body of the cells that were stained with PPM1H and PGM2L1 on each image (20–30 per image) and the mean gray value was obtained for each case. An approximately equal number of boxes was also drawn in the background of each image to measure the background mean gray values, which were also averaged over images. The background-subtracted mean intensity values of cellular profiles in the Huntington's disease samples were then compared to control samples using a two-tailed *t*-test. For anti-ID4, anti-HIBADH and anti-GFAP immunostainings, the number of ID4 or HIBADH positive cells with respect to GFAP positive cells was also analyzed: 8–15 images were also taken at 20× from the caudate of each control and Huntington's disease sample. The number of ID4-, HIBADH- and GFAP-positive cells on each image of each case was counted (the same number of images was used for each case across the three immunostainings). The count ratio of ID4 positive cells / GFAP positive cells and HIBADH positive cells / GFAP positive cells was calculated for each case, and Huntington's disease cases were compared to control cases using a two-tailed *t*-test.

15. Irizarry, R.A. *et al. Biostatistics* **4**, 249–264 (2003).
16. Leong, H.S., Yates, T., Wilson, C. & Miller, C.J. *Bioinformatics* **21**, 2552–2553 (2005).
17. Akaike, H. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
18. Burnham, K.P. & Anderson, D.R. *Model selection and multi-model inference* (Springer, 2002).
19. Cook, R.D. *J. Am. Stat. Assoc.* **74**, 169–174 (1979).
20. Benjamini, Y. & Hochberg, Y. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
21. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. *Bioinformatics* **20**, 307–315 (2004).
22. Steiner, P. *et al. Neuroscience* **113**, 893–905 (2002).
23. Levison, S. & McCarthy, K. in *Culturing Nerve Cells* (eds. Banker, G. and Goslin, K.) 309–335 (The MIT Press, 1991).

doi:10.1038/nmeth.1710