Blood and Brain Gene Expression Trajectories Mirror Neuropathology and Clinical Deterioration in Neurodegeneration

Yasser Iturria-Medina1,2*, Ahmed F. Khan1,2, Quadri Adewale1,2, Amir H. Shirazi1,2, the Alzheimer’s Disease Neuroimaging Initiative3.

1 McConnell Brain Imaging Center, Montreal Neurological Institute, Montreal, Canada
2 Ludmer Centre for NeuroInformatics and Mental Health, Montreal, Canada
3 Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

* Correspondence to: Yasser Iturria-Medina, 3801 University Street, room NW140, Montreal Neurological Institute, McGill University, Montreal, Canada H3A 2B4. Email: iturria.medina@gmail.com

Abstract

Most prevalent neurodegenerative disorders take decades to develop and their early detection is challenged by confounding non-pathological aging processes. For all neurodegenerative conditions, we continue to lack longitudinal gene expression (GE) data covering their large temporal evolution, which hinders the understanding of the underlying dynamic molecular mechanisms. Here, we overcome this key limitation by introducing a novel GE contrastive trajectory inference (GE-cTI) method that reveals enriched temporal patterns in a diseased population. Evaluated on 1969 subjects in the spectrum of late-onset Alzheimer’s and Huntington’s diseases (from ROSMAP, HBTRC and ADNI datasets), this unsupervised machine learning algorithm strongly predicts neuropathological severity (e.g. Braak, Amyloid and Vonsattel stages). Furthermore, when applied to in-vivo blood samples at baseline (ADNI), it significantly predicts clinical deterioration and conversion to advanced disease stages, supporting
the identification of a minimally invasive (blood-based) tool for early clinical screening. This technique also allows the discovery of genes and molecular pathways, in both peripheral and brain tissues, that are highly predictive of disease evolution. Eighty-five to ninety percent of the most predictive molecular pathways identified in the brain are also top predictors in the blood. These pathways support the importance of studying the peripheral-brain axis, providing further evidence for a key role of vascular structure/functioning and immune system response. The GE-cTI is a promising tool for revealing complex neuropathological mechanisms, with direct implications for implementing personalized dynamic treatments in neurology.

**Keywords:** gene expression trajectories, neurodegenerative progression, unsupervised machine learning, neuropathological mechanisms, personalized treatments.

**Introduction**

In recent decades, we have witnessed an accelerated characterization of the molecular and neuropathological mechanisms underlying neurodegenerative progression. Thanks to cutting-edge technological and methodological advances in genomic and proteomic analysis, we foresee unlimited methodological possibilities for understanding and modifying the role of genes and protein in disease (Esvelt and Wang, 2012; Mostafavi et al., 2018; Smith et al., 2016; Tan et al., 2012). Gene expression (GE) examination has been of crucial value, revealing disease-specific differentiated genes/molecular-pathways and gene-gene networks with a direct effect in neuropathological and cognitive/clinical deterioration (Mostafavi et al., 2018; Zhang et al., 2013). However, neurodegenerative conditions may take decades to develop and GE mapping techniques are quite recent, hence the unavailability of individual GE datasets covering a given disease’s whole evolution. All reported studies are based on cross-sectional or short-term longitudinal data, while we continue to lack long-term datasets covering the several phases underlying neurodegeneration.

In addition, due to its highly invasive nature, brain GE studies in neurodegeneration are based on *post-mortem* tissue samples. There are major challenges associated with the translation/extrapolation of *ex-vivo* results to *in-vivo* conditions (Ferreira et al., 2018). This could imply that disease mechanisms (e.g. gene-gene causal networks) and potential biomarkers
identified with *post-mortem* data may well not be entirely generalizable to live patients. In this sense, peripheral molecular measurements (e.g. plasma GE) may be used to cross-validate *post-mortem* based methodologies and findings, potentially providing minimally invasive *in-vivo* biomarkers for accurate patient screening in the daily clinic and clinical trials implementation. Nevertheless, the lack of comprehensive longitudinal peripheral datasets, covering multiple disease stages at the individual level, makes *in-vivo* dynamic molecular analyses unpractical. Consequently, this affects the identification of robust peripheral biomarkers across continuous disease stages and variants.

Due to the proven ability to disentangle temporal components from high-dimensional cross-sectional data, novel unsupervised Machine Learning (ML) techniques offer a viable opportunity for dealing with the previous limitations. The data-driven reconstruction of pseudo-temporal paths to order observations (e.g. cells, subjects) is revolutionizing *omics* studies, enabling for the first time the mapping of complex dynamic processes using cross-sectional “snapshots” (Cannoodt et al., 2016; Gupta and Bar-Joseph, 2008; Magwene et al., 2003; Welch et al., 2016). Based on the ML inference of a low dimensional space embedded in a population’s *omics* data, and by creating a relative ordering of the individuals, we can accurately identify a series of molecular states that constitute a longitudinal trajectory for a process of interest (Campbell and Yau, 2018). When used in RNA-seq studies, this novel technique has provided an unprecedented insight into the evolution of multiple pathologies. It has also allowed tracking and dissecting differentiated spatiotemporal programs in single-cell analysis (Briggs et al., 2018).

Driven by the imperative of a better understanding and an earlier detection of neurodegeneration, here we extend pseudotemporal trajectory inference (TI) to the analysis of both *post-mortem* and *in-vivo* (blood) GE neurodegenerative samples. Firstly, to better address important methodological limitations in data exploration and visualization, we introduce the contrastive Trajectory Inference (cTI) algorithm. This allows the unsupervised identification and ordering of enriched patterns in a diseased population (e.g. Alzheimer’s and Huntington’s diseases) relative to a comparison background population (e.g. healthy elderly). Next, we analyze GE samples from blood plasma of 744 subjects in the spectrum of late-onset Alzheimer’s disease (LOAD) and from 1225 autopsied brains in the spectrum of LOAD and Huntington’s disease (HD). Our method provides molecular pathological scores that are highly predictive of neuropathological and cognitive/clinical
deterioration. The results are strongly consistent for both in-vivo and post-mortem data. In addition, it allows identification of genes and molecular pathways driving neurodegenerative progression, as well as analysis of (dis)similarities in molecular disease mechanisms at brain and peripheral tissue levels. The inference of contrasted genetic trajectories is a promising tool for understanding complex neuropathological mechanisms and for minimally invasive patient screening at the daily clinic, with practical implications for implementing personalized medical interventions in neurology.

**Materials and Methods**

**Study Participants**

This study used GE data (N<sub>total</sub>=1969) from three large-scale databases (see Supplementary Table 1 for demographic characteristics). Each dataset was processed and analyzed independently:

**Dataset 1.** RNA expression data from the prefrontal cortex (PFC) of a subset of 489 autopsied subjects were downloaded from the Religious Orders Study (ROS; (Bennett et al., 2012a)) and the Memory and Aging Project Study (MAP; (Bennett et al., 2012b)). This data (Bennett et al., 2018) is available at the Accelerating Medicines Partnership Alzheimer’s Disease (AMP-AD) knowledge portal (https://www.synapse.org/#, Synapse ID 3800853). ROS (Bennett et al., 2012a) and MAP (Bennett et al., 2012b) are longitudinal clinical-pathologic cohort studies of aging, Alzheimer's disease (AD) and related disorders. Enrollment required no known sign of dementia. Upon death, a post-mortem neuropathologic evaluation is performed that includes a uniform structured assessment of AD pathology, cerebral infarcts, Lewy body disease, and other pathologies common in aging and dementia. The pathologic diagnosis of AD uses NIA-Reagan and modified CERAD criteria, and the staging of neurofibrillary pathology uses Braak Staging (Braak H, 1991). An RNA integrity (RIN) score >5 and a quantity threshold (5 mg) for each sample were required (Bennett et al., 2014). cRNA was hybridized to Illumina HT-12 Expression Bead Chip (48,803 transcripts) via standard protocols using an Illumina Bead Station 500GX (Webster et al., 2009; Zhang et al., 2013).

**Dataset 2.** 736 individual post-mortem tissue samples from the dorsolateral prefrontal cortex BA9 of LOAD patients (N=376), HD patients (N=184) and nondemented subjects (N=173) were
collected and analyzed (Zhang et al., 2013). All autopsied brains were collected by the Harvard Brain Tissue Resource Center (HBTRC; GEO accession number GSE44772), and include subjects for whom both the donor and the next of kin had completed the HBTRC informed consent (http://www.brainbank.mclean.org/). Correspondingly, tissue collection and the research were conducted according to the HBTRC guidelines (http://www.brainbank.mclean.org/). Postmortem interval (PMI) was 17.8 ± 8.3 hr, sample pH was 6.4 ± 0.3 and RNA integrity number (RIN) was 6.8 ± 0.8 for the average sample in the overall cohort.

As previously described in (Zhang et al., 2013), RNA preparation and array hybridizations applied custom microarrays manufactured by Agilent Technologies consisting of 4,720 control probes and 39,579 probes targeting transcripts representing 25,242 known and 14,337 predicted genes. Arrays were quantified on the basis of spot intensity relative to background, adjusted for experimental variation between arrays using average intensity over multiple channels, and fitted to an error model to determine significance (Emilsisson et al., 2008). Braak stage, general and regional atrophy, gray and white matter atrophy and ventricular enlargement were assessed and cataloged by pathologists at McLean Hospital (Belmont, MA, USA). In addition, the severity of pathology in the HD brains was determined using the Vonsattel grading system (Vonsattel et al., 1985).

Dataset 3. This study used a total of 744 individual data with blood GE information, from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu). The participants underwent multimodal brain imaging evaluations, including amyloid PET, tau PET and/or structural MRI. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD).

The Affymetrix Human Genome U219 Array (www.affymetrix.com) was used for gene expression profiling from blood samples. Peripheral blood samples were collected using PAXgene tubes for RNA analysis (Saykin et al., 2015). The quality-controlled GE data includes activity levels for 49,293 transcripts. All the participants were characterized cognitively using the mini-mental state examination (MMSE), a composite score of executive function (EF), a composite score of memory integrity (MEM) (Gibbons et al., 2012), and Alzheimer's Disease Assessment Scale-Cognitive
Subscales 11 and 13 (ADAS-11 and ADAS-13, respectively). Also, they were clinically diagnosed at baseline as healthy control (HC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) or probable Alzheimer’s disease patient (LOAD).

$^{18}$F-AV-45 (amyloid specific) and $^{18}$F-AV-1451 (tau specific) PET images were acquired for a subset of 660 and 166 subjects, respectively. Both amyloid and tau images were preprocessed by the Jagust Lab (UC Berkeley, US; Jagust et al., 2010). Using the amyloid images, subjects were categorized as amyloid positive (Aβ+) or negative (Aβ-) by applying a cutoff of 1.11 to a Florbetapir composite SUVR normalized by the whole cerebellum reference (Described in ADNI_UCBERKELEY_AV45_Methods_12.03.15.pdf file, ADNI database). Also, individual Freesurfer-defined cortical and subcortical brain regions were used to calculate weighted Flortaucipir averages for each region, which were normalized by the weighted Flortaucipir at the cerebellum (Described in UCBERKELEY_AV1451_Methods_Aug2018.pdf file, ADNI database). Based on the lobar classification topographic staging scheme for tau PET and the corresponding cutoff values proposed by (Schwarz et al., 2018), the subjects were staged in Braak 0 (no tau), Braak I/II, Braak III/IV or Braak V/VI. Subsequently, they were categorized as tau negative (tau-) or positive (tau+) if they were in the stages 0 or I-VI, respectively. Structural MRI images for 741 subjects were analyzed by a physician specially trained in the detection of MRI infarcts. The presence of MRI infarction was determined from the size, location and imaging characteristics of the lesion, with only lesions 3mm or larger qualifying for consideration as cerebral infarcts (Described in ADNI_UCD_MRI_Infarct_Assessment_Method_201130609.pdf file, ADNI database). Finally, a subset of subjects (N=30) was evaluated for pathological brain lesions after death. Pathological lesions were assessed using established neuropathologic diagnostic criteria (Described in ADNI_Methods_Neuropathology_Core_03-06-2018-2.pdf file, ADNI database). The analysis included histopathologic assessments of amyloid β deposits, staging of neurofibrillary tangles, scoring of neuritic plaques and assessments of co-morbid conditions such as Lewy body disease, vascular brain injury, hippocampal sclerosis, and TAR DNA binding protein (TDP) immunoreactive inclusions (Montine et al., 2012).

**Contrastive Trajectories Inference (cTI)**
Given a multi-dimensional population dataset, the inference of contrasted pseudotemporal trajectories (and an individual pseudotime value) consists of four main steps (see also Code availability):

(i) For high-dimensional datasets (e.g. ~40,000 transcripts), initial selection of features most likely to be involved in a trajectory across the entire population. We apply the unsupervised method proposed by (Welch et al., 2016), which does not require prior knowledge of features involved in the process or differential expression analysis. Features are scored by comparing sample variance and “neighborhood variance”. Specifically, for a gene transcript \( g \), its sample variance \( \sigma_g^2 \) across all samples is calculated. Then, the “neighborhood variance” is computed as:

\[
S_g^{2(N)} = \frac{1}{N_{\text{transcripts}} k_c - 1} \cdot \sum_{i=1}^{N_{\text{genes}}} \sum_{j=1}^{k_c} (e_{ij} - e_{N(i)g})^2,
\]

where \( N_{\text{transcripts}} \) is the total number of gene transcripts, \( e_{ij} \) is the expression level of the \( j^{\text{th}} \) transcript in the \( i^{\text{th}} \) sample, \( N(i, j) \) is the \( j^{\text{th}} \) nearest neighbor of sample \( i \), and \( k_c \) is the minimum number of neighbors needed to yield a connected graph. \( S_g^{2(N)} \) is similar to the sample variance computed with respect to neighboring points rather than the mean, measuring how much \( g \) varies across neighboring samples. Intuitively, gene transcripts most likely to be involved in a trajectory should present a more gradual variation across neighboring points than at global scale, which would correspond to a high ratio \( \sigma_g^2 / S_g^{2(N)} \). Thus, a threshold is applied to select those features with higher \( \sigma_g^2 / S_g^{2(N)} \) score, e.g. we kept the features with at least a 0.95 probability of being involved in a trajectory (i.e. ~3000 gene transcripts).

(ii) Data exploration and visualization via contrastive Principal Component Analysis (cPCA; (Abid et al., 2018)). This novel technique identifies low-dimensional patterns that are enriched in a target dataset (e.g. a diseased population) relative to a comparison background dataset (e.g. demographically matched healthy subjects). By controlling the effects of characteristic patterns in the background (e.g. pathology-free and spurious associations, noise), cPCA (Abid et al., 2018) allows visualizing specific data structures missed by standard data exploration and visualization methods (e.g. PCA, Kernel PCA). Specifically, if \( C_{\text{target}} \) and \( C_{\text{background}} \) are the covariance matrices of the target and background data, the directions returned by cPCA are the singular vectors of the weighted difference of the covariance matrices: \( C_{\text{target}} - \alpha C_{\text{background}} \). The
contrast parameter $\alpha$ represents the trade-off between having the high target variance and the low background variance. Multiple values of $\alpha$ are used (i.e. 100 logarithmically equally spaced points between $10^{-2}$ and $10^2$). Instead of choosing a single $\alpha$, the resulting subspaces for all the $\alpha$ values are clustered (based in their proximity in terms of the principal angle and spectral clustering (Ng et al., 2002)) in a few subspaces. The data is then projected onto each of these few subspaces, revealing different trends within the target data. While the original cPCA algorithm proposes to select the final subspace via visual examination, we chose automatically the subspace that maximize the clustering tendency in the projected target data. For this, the ‘gap’ cluster evaluation criterion, implemented in the MATLAB function evalclusters, is employed.

When applied cPCA to the selected GE transcripts (from step i), for each population, we obtained around 6-8 contrasted principal components capturing the most enriched pathological properties relative to the background (i.e. subjects without cognitive deterioration and neuropathological signs). For ROSMAP, HBTRC and ADNI, sample sizes of the background populations were 177 (36%), 173 (23%) and 113 (15%), respectively. Selected $\alpha$ values for these three studied datasets were: 11.76 (ROSMAP), 17.07 (HBTRC) and 11.76 (ADNI).

(iii) Subjects ordering and GE-pseudotime calculation according to their proximity to the background population in the contrasted Principal Components space. For this, we first calculate the Euclidean Distance Matrix among all the subjects and the associated Minimum Spanning Tree (MST). The MST is then used to calculate the shortest trajectory/path from any subject to the background subjects. Each specific trajectory consists of the concatenation of relatively similar subjects, with a given behavior in the data’s dimensionally reduced space. The position of each subject in his/her corresponding shortest trajectory reflects the individual proximity to the pathology-free state (the background) and, if analyzed in the inverse direction, to advanced disease state. Thus, to quantify the distance to these two extremes (background or disease), an individual GE-pseudotime score is calculated as the shortest distance value to the background’s centroid, relative to the maximum population value (i.e. values are standardized between 0 and 1). Finally, the subjects are ordered according to their GE-pseudotime values, from low (close to the background group) to high values (close to the most diseased subjects).

Additionally, in order to evaluate cPCA’s performance versus other popular dimensionality reduction techniques, we repeated step (ii) using the traditional PCA (Abdi and Williams, 2010).
and the recently proposed non-linear Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) approach (McInnes and Healy, 2018). Subsequently, we re-applied step (iii), obtaining alternative subjects orderings (and GE-pseudotimes) according to their proximity to the background population in the resulting PCA and UMAP Components space.

Statistics

Data preprocessing: Before applying the cTI approach, each gene transcript’s activity was adjusted for relevant covariates using robust additive linear models (Street et al., 1988). Specifically, Dataset 1 GE was adjusted for postmortem interval (PMI) in hours, age, gender and educational level. Dataset 2 GE was adjusted for PMI, sample pH, RNA integrity number (RIN), age and gender. Dataset 3 GE was controlled for RIN, Plate Number, age, gender and educational level. Also, each adjusted gene transcript activity was approximately transformed into a normal distribution via the box-cox transformation (Box and Cox, 1964), implemented in the MATLAB function boxcox.

Post-hoc analyses: All predictive associations between grouping variables (e.g. Braak, Cerad and Vonsattel stages, clinical diagnosis) and the individual GE-pseudotimes (see first and second Results subsections) were tested via ANOVA tests, FWE-controlled by permutations (Legendre and Legendre, 1998). For each dataset, the total contribution \( C_i \) of each gene transcript \( i \) to the obtained reduced representation space (and the genetic trajectories) was quantified as (Abdi and Williams, 2010):

\[
C_i = 100 \cdot \sum_{j=1}^{N_{cPC}} \left( \lambda_j^{norm} \cdot \frac{\omega_{i,j}^2}{\sum_{k=1}^{N_{genes}} \omega_{i,k}^2} \right),
\]

where \( \lambda_j^{norm} = (\lambda_j - \text{min}_\lambda) / \sum_{k=1}^{N_{total}} (\lambda_k - \text{min}_\lambda) \) is the normalized eigenvalue of the contrasted principal component \( j \), \( \text{min}_\lambda \) is the minimum obtained eigenvalue, \( N_{total} \) is the original number of contrasted principal components, \( N_{cPC} \) is the number of contrasted principal components with \( \lambda_j^{norm} \) over a predefined cut-off value (i.e. 0.025), \( \omega_{i,j} \) is the loading/weight of the gene transcript \( i \) on the component \( j \), and \( N_{features} \) is the total number of gene transcripts considered in the dimensionality reduction analysis. Similarly, the expected contribution value (cut-off) was calculated as (Abdi and Williams, 2010):
\[ C_{\text{expected}} = 100 \cdot \sum_{j=1}^{N_{\text{EPC}}} \left( \lambda_j^{\text{norm}} \cdot \frac{1}{N_{\text{features}}} \right). \] 

The gene transcripts with total contribution \( C_i \) over the expected contribution value \( C_{\text{expected}} \) were considered as highly influential to obtain the reduced representation space.

**Results**

*Inferring Enriched GE Neurodegenerative Trajectories*

GE, neuropathology and cognitive/clinical deterioration in 1969 demented and non-demented subjects from three large-scale studies were assessed (see Figure 1, and Datasets 1-3 in *Materials and Methods*). GE and neuropathology evaluations from both dataset 1 (N=489, R OSMAP Study) and dataset 2 (N=736, HBTRC database) were performed in autopsied brains, with genetic profiling from the PFC. GE from Dataset 3 (N=744, ADNI database) was obtained from *in-vivo* blood samples, with all subjects also having brain imaging evaluations including amyloid PET, tau PET and/or structural MRI.

Aiming to uncover the molecular reconfigurations underlying neurodegenerative evolution, we proceeded to reorder the GE patterns (Fig. 1). For this, we implemented a novel unsupervised algorithm for detecting enriched trajectories in a diseased population relative to a background dataset (e.g. normal controls; see cTI subsection in *Materials and Methods*). A distinctive feature of cTI is the use of a contrastive PCA algorithm (Abid et al., 2018), which controls by the principal components of the background data to optimize the exploration and visualization of the target. It is a generic algorithm, adaptable to different types of data (e.g. genomic, proteomic, imaging, clinical). Each GE dataset was firstly adjusted for relevant confounding covariates (e.g. RIN, age, gender and/or educational level; see details in *Statistics, Materials and Methods*). Next, the cTI was independently applied to the three populations, providing population-specific trajectories starting on the background data. Each trajectory is composed by the concatenation of a subset of subjects, which follows a given behavior in the data’s dimensionally reduced space. We hypothesized that the position of each subject in these GE trajectories would reflect individual proximity to the pathology-free state (the background) or, if analyzed in the inverse direction, proximity to advanced disease states. Correspondingly, a GE-pseudotime value ([0,1] range) is
calculated for each subject, with relatively low values for subjects with final positions close to the background data, and high values for subjects on the distant extremes of the population. Notice that GE-pseudotime could then be assumed as an individual molecular score of pathological progression, whose validity is tested in the following subsections. See also Figure 1.

Figure 1. Schematic approach for GE-trajectories analysis in neurodegeneration. a) *In-vivo* blood (N=744) and *post-mortem* brain (N=1225) tissues collected. b) RNA expression for around 40,000 transcripts (dataset-specific). c) The high dimensional data is automatically reduced to an enriched space (~5 features) via a contrastive PCA algorithm (cPCA (Abid et al., 2018)), which optimizes the exploration and visualization of the target population’s data. d) In the contrasted Principal Components (cPC) space, each subject is assigned to a GE-trajectory. The subject’s position in the corresponding GE-trajectory reflects the individual proximity to the pathology-free state (the
background) and, if analyzed in the inverse direction, to the advanced disease state. An individual GE-pseudotime score is calculated, reflecting the distance to these two extremes (background or disease). e) When taken as an individual molecular score of disease evolution, the GE-pseudotime significantly associate with neuropathological and/or cognitive measurements. f) Both in peripheral and brain tissues, the cPCA’s loadings (or weights) allow the identification and posterior functional analysis of most informative genes in terms of pathological evolution.

**Post-mortem GE Trajectories Predict Neurodegenerative Severity**

Firstly, we analyzed the GE trajectories obtained for the ROSMAP study (dataset 1, N=489). The results (Figs. 2a-c) showed a clear association between the obtained molecular disease score (GE-pseudotime) and the autopsied tau and amyloid assessments, with a higher GE-pseudotime value implying an advanced neuropathologic state. Group differences in GE-pseudotime values were statistically tested via ANOVA tests with permutations. We found robust significant associations between the GE-pseudotimes and Braak stages (Fig. 2a; F=4.09, P=0.001, FWE-corrected), Cerad stages (Fig. 2b; F=9.23, P<0.001, FWE-corrected), and a composite variable (Braak+Cerad) reflecting the simultaneous presence of tau and amyloid (Fig. 2c; F=5.97, P<0.001, FWE-corrected).

Next, we explored the generalizability of these results in the considerably more heterogeneous database from HBTRC (dataset 2, N=736), including two different disorders (LOAD and HD) and nondemented controls. In consistence with the previous findings, we observed (Figs. 2d-e) a positive association between the individual molecular disease score and the levels of neuropathologic affectation in both disorders. The GE-pseudotimes were significantly associated with the Braak stages (Fig. 2d; F=11.17, P<0.001, FWE-corrected) and the Vonsattel stages (Fig. 2e; F= 9.04, P<0.001, FWE-corrected). The fact that this population included multiple disorders did not seem to affect the robustness of the subject ordering in relation with disease progression, which supports the identification of a promising biomarker for the analysis of comorbid neurological conditions.

Importantly, when compared the individual molecular disease scores obtained using the contrastive PCA algorithm (Abid et al., 2018) with the obtained using the traditional PCA and the
novel non-linear UMAP approaches (see Methods, Contrastive Trajectories Inference subsection), we observed that cPCA-based results significantly outperformed PCA- and UMAP-based results. Essentially, the GE-pseudotimes obtained with PCA- and UMAP didn’t show any significant association with neuropathological variables (all $P>0.3$, FWE-corrected; see Fig. S1). This finding strongly supports the key advantage of considering the enriched patterns in the population of interest relative to the background dataset (Abid et al., 2018).
Figure 2. GE-based predictions of neurodegenerative severity for ROSMAP, HBTRC and ADNI populations. a-e) GE-pseudotime predictive associations with Braak (a,d), Cerad (b), Braak for Aβ-/Aβ+ (c), and Vonsattel (e) stages in ROSMAP (a-e) and HBTRC (d-e). f-j) GE-pseudotime predictive associations with Tau positivity (f), Aβ positivity (g), Tau-Aβ comorbidity (h), cerebral infarcts occurrence (i) and Tau-Aβ-infarcts comorbidity (j) in ADNI population. Data is GE-
pseudotime mean (± standard error). All P values are FWE-corrected (see reported values in Results).

**Blood GE as a Robust Biomarker of In-Vivo Neuropathological Severity and Clinical Deterioration**

Next, we aimed to investigate if the unsupervised ordering of GE patterns present in the blood can reflect neuropathological severity and, importantly, if it could be used as a marker of present and future clinical deterioration. If successful, the latter could have strong implications for the *in-vivo* detection of future disease evolution in the clinic and to decide if a patient should be therapeutically treated or not. To test this, we identified the enriched GE trajectories in the plasma of 744 participants in the spectrum of LOAD from ADNI (dataset 3), taking as background 113 subjects without cognitive/clinical alterations or any evidence of amyloid deposition or cerebral infarcts (see amyloid PET imaging and neuropathology evaluation for Dataset 3, in Materials and Methods).

In line with our previous findings with the ROSMAP and HBTRC *post-mortem* data, the ADNI-based results (Figs. 2f-j) showed a significant predictive power of pathological severity. The individual GE-pseudotime values vastly reflected the differences in tau positivity (Fig. 2f; \(F=17.64, P<0.001\), FWE-corrected), amyloid positivity (Fig. 2g; \(F=28.22, P<0.001\), FWE-corrected), tau-amyloid comorbidity (Fig. 2h; \(F=9.58, P<0.001\), FWE-corrected), brain infarcts (Fig. 2i; \(F=5.32, P<0.05\), FWE-corrected), and tau-amyloid-infarcts comorbidity (Fig. 2j; \(F=7.49, P<0.001\), FWE-corrected).

In addition, we tested if the identified subject ordering based on enriched GE patterns was predictive of the individual clinical and cognitive properties (Figs. 3a-d). We observed that the molecular disease score values were significantly associated with the individual clinical diagnosis (Fig. 3a; \(F=56.72, P<0.001\), FWE-corrected). Importantly, they were also significantly associated with the individual clinical conversion (Fig. 3b; \(F=56.61, P<0.001\), FWE-corrected). Subjects with a same clinical diagnosis at baseline, but significantly higher GE-pseudotimes, were consistently progressing to an advancer disease state in an average period of 3.18 years (std: 2.33). The molecular disease score values were also significantly associated with executive function (Fig.
However, the associations with these continuous cognitive metrics were characterized by a low predictive power, only explaining around 5.3% to 7.3% of the population variance, respectively. We attribute this to both the lack of highly precise metrics for evaluating memory and executive function and the inability of the GE-pseudotimes to reflect specific aspects/components of each individual’s cognitive deterioration.

Altogether, these results support that, in the context of in-vivo LOAD and the ADNI population, the subject’s temporal ordering based on enriched blood GE patterns is strongly reflective of neuropathological and overall clinical deterioration, as well as future disease progression. It is, however, a considerably less powerful predictor of the detailed alterations observed in memory and executive function.
Figure 3. Blood GE-based predictions of Clinical and Cognitive Deterioration for ADNI data. 
a-b) GE-pseudotime associations with clinical diagnosis (a) and future clinical conversion (b). 
c-d) Scatter plots showing negative associations between molecular disease progression (reflected in 
GE-pseudotime) and measurements of cognitive integrity: MEM (c) and EF (d). In (a-b), data is 
GE-pseudotime mean (± standard error) and P values are FWE-corrected. In (b), categories 
included are: stable HC (s-HC), converter HC (c-HC), stable MCI (s-MCI) and converter MCI (c-
MCI).

In-Vivo and Post-Mortem Molecular Pathways Underlying LOAD Progression
Next, we aimed to identify the genes, molecular functions and pathways responsible for the accurate prediction of neurodegenerative progression in LOAD. We also intended to clarify if similar predictive mechanisms were common to the periphery (blood) and brain tissues. In this context, the GE-cTI can provide a quantitative mapping of the most influential genes during the process of diseased trajectories inference. Specifically, the cPCA’s loadings (or weights) reflect how much each specific gene, in the original high dimensional space (i.e. ~40K transcripts), contributed to the reduced low dimensional space from which the trajectories were obtained. Thus, we used these weights to select the genes most influential on the subject’s ordering, i.e. those genes driving the observed population differences predictive of neuropathological and cognitive/clinical alterations across the disease’s evolution (see Statistical Analysis, Materials and Methods). Based on the dataset-specific identified genes, we then performed large-scale gene functional analyses with the Protein Annotation Through Evolutionary Relationship (PANTHER) classification system (Mi et al., 2013). In addition, using a recently reported comprehensive meta-analysis of brain cell-type gene signatures (Mckenzie et al., 2018), we identified the cell-types consistently associated with the most predictive genes in the brain. Of note, because these analyses were restricted to LOAD evolution, HD patients were excluded when using the HBTRC database.

For the ROSMAP brains, we found 845 highly influential genes with 88 functional pathways (Figs. 4a,e, and Supplementary Tables 2,3). These GO overrepresented pathways were highly sensitive for the detection of biological processes that are commonly associated with neuropathological and cognitive deterioration mechanisms, including axon guidance, histamine H1 receptor mediation, angiogenesis, inflammation mediated by chemokine and cytokine signaling, Wnt and VEGF signaling, apoptosis, p53 pathway, and Alzheimer’s disease-amyloid secretase. For HBTRC brains, we found 416 highly influential genes with 74 functional pathways (Figs. 4b,e, and Supplementary Tables 2,4). Eighty-nine percent of these pathways (i.e. 66) were also among the most relevant pathways detected in ROSMAP brains. Correspondingly, the GO overrepresented pathways in HBTRC brains were also highly sensitive for the detection of biological processes commonly associated with neurodegeneration.

The highly predictive genes in both ROSMAP and HBTRC were consistently related to five different cell-types (Fig. 4d), including astrocytes, endothelial cells, microglia, neurons, and oligodendrocyte precursor cells. Interestingly, the patterns of identified cell-types differed between
these two brain datasets (Fig. 4d). Astrocytes (38%) and oligodendrocyte precursor cells (35%) were the most abundant identified cell-types in ROSMAP, while microglia cells were underrepresented (4%). Contrary, almost half of the identified cell-types in HBTRC corresponded to microglia (48%), although astrocytes and oligodendrocyte precursor cells still represented significant proportions (22% and 17%, respectively). As discussed below, the observed inter-dataset differences in molecular pathways and brain cell-types may be responding to multiple causes, such as the systematic (study-specific) sampling at distinct brain locations, the use of different GE mapping techniques with dissimilar sensitivity/specificity capacities, and different population characteristics.

Notably, 85% and 90% of the highly predictive molecular pathways in the neurodegenerating brain (ROSMAP and HBTRC, respectively) were also among the most relevant pathways detected in the blood data (ADNI; Figs. 4c,e, and Supplementary Tables 2,5). The common blood-brain functional pathways relevant for LOAD progression included blood coagulation, angiogenesis (linked to the formation of new blood vessels), p53 (modulating the cell cycle and playing a major role in inhibition of angiogenesis), B cell activation (involved in immune system response), Wnt signaling (related to signal transduction), among others (Fig. 4e and Supplementary Tables 3-5).

The finding of these common pathways evidence the direct relationship between the central nervous system and the body, both in health and in disease. Their unsupervised data-driven identification is, therefore, supporting the crucial importance of studying the periphery-brain axis (e.g. immune and vascular interactions with brain integrity) for a better understanding of systemic pathological mechanisms underlying neurodegeneration.

Interestingly, we also found another 15% and 10% of highly predictive molecular pathways in the blood that were not identified in the brain (ROSMAP and HBTRC, respectively; Fig. 4e and Supplementary Tables 3-5). Similarly, 11% of the most predictive pathways identified in HBTRC were not common with the pathways identified in ROSMAP, and a clear difference in each pathway’s presence level across the three datasets was also noticed (Fig. 4e and Supplementary Tables 3-5). These findings may be associated with several reasons, including increased pathological comorbidity in the periphery relative to the brain and/or crucial methodological limitations, such as the analysis of three different populations with divergent disease
characteristics, and the use of different GE mapping techniques with dissimilar sensitivity/specificity capacities.

Figure 4. Ontology analysis of top predictor genes of LOAD development. Significant molecular functions (a,b,c), cell-types (d) and molecular pathways (e) identified in the brain’s PFC (ROSMAP) and BA9 (HBTRC) areas, and the blood plasma (ADNI). In (a-c), the bars show the number of genes associated to the main GO overrepresented categories. In (e), only common pathways identified across the three populations are shown (for complete lists of genes and pathways, see Supplementary Tables 2-5). The color scale indicates the presence level (in %) of each functional pathway (e.g. dark blue for absent pathways, red for highly represented pathways).

Discussion
Due to the typically long developing period of most prevalent neurodegenerative disorders, we lack exhaustive longitudinal datasets covering the continuous molecular transitions underlying disease progression. Consequently, almost all our knowledge of the subjacent pathological mechanisms is based on data “snapshots” taken and analyzed at a few disease stages. Here, we aimed to overcome this crucial gap by inferring the intrinsic temporal information contained in large-scale neurodegenerative datasets. For that, we implemented a novel pattern analysis method that detects enriched GE trajectories in a diseased population (e.g. subjects progressing towards dementia) relative to a background population (e.g. a clinically normal control group). Our results in three different GE datasets (ROSMAP, HBTRC, ADNI) support the strong predictive power of this technique for identifying individual neuropathological stages and/or cognitive deterioration. This may well have broad implications for uncovering the dynamic mechanisms of molecular pathology, patient stratification in the clinic, and monitoring response to personalized treatments in neurodegeneration.

A minimally invasive molecular test for neurodegeneration could lead to better treatment and therapies (Park et al., 2019; Ray et al., 2007). An additional aim of this study was to identify an in-vivo peripheral biomarker able to predict the individual’s pathophysiology and cognitive decline. When tested in 744 blood samples from ADNI, the proposed GE-cTI showed a significant association with amyloid, tau and infarcts positivity (Figs. 2f-j). Furthermore, it was significantly associated with clinical deterioration and conversion (Figs. 3a-b). The fact that the proposed ML model is un-supervised (i.e. the neuropathological and clinical variables are not used to train a predictive model), guarantees absence of possible circularity or data overfitting. Consequently, we can infer that the obtained genetic trajectories and the associated GE-pseudotime values are direct measures of molecular integrity, obtained independently of phenotypic variables, and would therefore be useful as unbiased biomarkers in clinical applications.

Our analysis of most relevant molecular pathways for predicting LOAD progression revealed a striking similarity between peripheral and intra-brain pathological mechanisms. Eighty five to ninety percent of the most predictive molecular pathways identified in the post-mortem brains were also identified as top predictors in the blood. These pathways support the importance of studying the peripheral-brain axis (see Results, last subsection), providing further evidence for a key role of vascular structure and functioning (Bell and Zlokovic, 2009; Iturria-Medina et al.,
2017, 2016), and immune system response (Gendelman, 2002; Labzin et al., 2018; Streit et al., 2004). The multi-tissue analysis based on genetic trajectories may be particularly useful for clarifying both local (tissue-specific) and systemic (inter-organs) neurodegenerative mechanisms.

Our method built on the pseudotemporal trajectory inference field (Cannoodt et al., 2016; Gupta and Bar-Joseph, 2008; Magwene et al., 2003; Welch et al., 2016). Modeling the dynamics of gene regulation, rather than focusing on static time points, is crucial for clarifying cellular transitions and what goes wrong in the case of disease (Cannoodt et al., 2016). We attempted to extend previous models by incorporating the use of a novel contrastive dimensionality reduction technique (cPCA; Abid et al., 2018), which allows detecting enriched patterns in the population of interest while adjusting by confounding components in the background population (e.g. concurrent aging effects). We observed that this technique (cPCA) was significantly more sensitive to detecting disease progression than other popular dimensionality reduction methods (i.e. PCA and UMAP; see Results and Fig. S1) In a set of complementary analyses (data not shown), we observed that, in comparison with other state-of-the-art TI methods (Campbell and Yau, 2018; Welch et al., 2016), this extension provides a considerably higher sensitivity to detect diseased GE components (i.e. other methods could not predict neuropathology, nor clinical deterioration). In addition to uncovering disease dynamics, cTI may enable the data-driven identification of new subpopulations within a heterogeneous neurodegenerative population (Cannoodt et al., 2016; Trapnell, 2015; Trapnell et al., 2014), with strong implications for precision medicine and the selective enrollment of patients in clinical trials. Furthermore, once the data is ordered, it could also improve the inference of causative regulatory interactions underlying a disorder (Cannoodt et al., 2016).

Another advantage of cTI (and TI in general) is the ability to deal with high dimensional data. This is a key feature for the concurrent analysis of multi-omics, potentially allowing the exploration of multiple and complementary modalities, such as transcriptomics, proteomics, metabolomics and epigenomics. Contrastive trajectory inference can be also applied to the analysis of data from other fields, including multi-modal brain imaging, environmental and cognitive/clinical information. Finally, although our study is focused on neurodegenerative evolution, in general, cTI can be applicable to the study of multiple neurological and neuropsychiatric conditions.
Data and Code availability

The three datasets used in this study are available at the AMP-AD knowledge portal (https://www.synapse.org/#, Synapse ID 3800853), the Gene Expression Omnibus (GEO accession number GSE44772) and the ADNI database (www.adni.loni.usc.edu), respectively. We anticipate that the cTI method will be released soon as part of an open-access user-friendly software. In the meantime, the MATLAB codes can be downloaded from our lab http://www.neuropm-lab.com.

Acknowledgments

This research was undertaken thanks in part to funding from: the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives Initiative, and the Brain Canada Foundation and Health Canada support to the McConnell Brain Imaging Center at the Montreal Neurological Institute. We thank the two anonymous reviewers whose comments/suggestions helped improve and clarify this manuscript. Also, Dataset-1 (ROSMAP) was provided by the Rush Alzheimer’s Disease Center, Rush University Medical Center, Chicago. Dataset-1 collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, U01AG46152, and the Illinois Department of Public Health. ROSMAP data can be requested at www.radc.rush.edu. In addition, dataset-3 collection and sharing for this project was funded by ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda
Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Author Contributions

ROSMAP, HBTRC and ADNI acquired the data. YIM conceived the study, implemented the programming source code, preprocessed and analyzed the data, and wrote the draft manuscript. AFK and QA prepared the figures and corrected the manuscript. AHS corrected the cPCA code. All authors contributed to constructive discussions.

References


Religious Orders Study and Rush Memory and Aging Project. J. Alzheimer’s Dis. 64, S161–S189. doi:10.3233/JAD-179939


McInnes, L., Healy, J., 2018. UMAP: Uniform Manifold Approximation and Projection for


Webster, J.A., Gibbs, J.R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A.,
