SUPPLEMENTARY FILE

The Intracerebral Hemorrhage Blood Transcriptome in Humans Differs from the Ischemic Stroke Blood Transcriptome

Boryana Stamova[1], Bradley P Ander[1], Glen Jickling[1,2], Farah Hamade[1], Marc Durocher[1], Xinhua Zhan[1], DaZhi Liu[1], Xiyuan Cheng[1], Heather Hull[1], Alan Yee[1], Kwan Ng[1], Natasha Shroff[1], Frank R Sharp[1]

[1]University of California at Davis, Sacramento, United States, [2]University of Alberta, Edmonton, Canada

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Angiogenesis and Vasculogenesis Genes with Differentially Expressed Transcripts in ICH vs. CTRL (A) and IS vs. CTRL (B). Green denotes down-regulation and red denotes up-regulation of the genes. Orange color in the Angiogenesis/Vasculogenesis process in ICH – positive Z score for predicted activation (not passing the threshold of Z>2 for significance though). Blue color for Angiogenesis in IS – negative Z score for predicted inhibition of the process (not passing the threshold of Z<-2 for significance though).

Figure S2. Transcriptome architecture of the 63 DET in IS as compared to CTRL. Biotypes of the DET.

Figure S3. Predicted Upstream Regulators of the Observed Transcriptome Changes Following ICH and IS. * Significant Z-score and significant p-value ($p<0.05$). PSEN1 itself was downregulated in IS vs. CTRL (FC=-1.7) and TGFA itself was up-regulated in ICH vs. CTRL (FC=1.9). Yellow highlighted gene symbols – $p<0.01$ for prediction of an upstream regulator in ICH vs CTRL and/or in IS vs CTRL. Not highlighted gene symbols – $0.01<p<0.05$.

Figure S4. Transcriptome architecture of ICH as compared to IS. Biotypes of the DET. Bright orange arrow points to T-Cell receptor genes, light grey arrow – to non-coding RNA (ncRNA).

Figure S5. Over-Represented Canonical Pathways (A, top 25 pathways) and Differentially Expressed Transcripts from Cell Type - Specific Genes (B) in ICH vs IS. A. The X-axis represents negative $\log_{10}$ (Benjamini-Hochberg corrected p-value). $-\log_{10}$ (Benjamini-Hochberg corrected p-value) >1.3 corresponds to Benjamini-Hochberg corrected p-value < 0.05. Orange bars represent predicted activation of the pathway, blue bars – predicted inhibition of the pathways, grey bars – direction cannot be predicted. Pathways having significant Z-scores for activation (Z>2) or for inhibition (Z<-2) are denoted to the right of the pathway bar. B. DET of cell-type specific genes in the 256 DET from ICH vs IS analysis. P-values represents the

hypergeometric probability of overlap between our DET list (by gene symbol) and the list of cell-type specific genes from the HaemAtlas[1]. Red – genes with upregulated DET (expressed higher in ICH than in IS); green – genes with down-regulated DET (expressed lower in ICH than in IS).

Figure S6. Transcriptome Architecture During Day 1 (within 24 hours post event) following IS and ICH (A). Biotypes of the DET. Asterisks denote significant difference between the ICH and IS transcriptomes at Day 1 for the particular transcript biotype. Bright orange arrow points to T-Cell receptor genes, light grey arrow – to non-coding RNA (ncRNA). B. Transcript-level overlap of IS (TP1) vs CTRL (TP0) and ICH (TP1) vs CTRL (TP0). Venn diagrams represent the number of DET. Red – genes with upregulated DET (expressed higher in ICH than in CTRL); green – genes with down-regulated DET (expressed lower in ICH than in CTRL).

Figure S7. Principal Components Analysis on the differentially expressed transcripts (DET) between controls subjects (CTRL, TP0) and: A. Intracerebral Hemorrhage (ICH) patients within 24 hours post-onset (ICH_TP1), 2667 DET. B. Ischemic Stroke (IS) patients within 24 hours post-onset (IS_TP1), 311 DET. On Scan Date-, Age- and Sex-effect corrected data.

Figure S8. Time-dependent profiles as determined by SOM (Self-Organizing Maps). Transcriptome architecture of ICH and IS as compared to CTRL. The biotypes of RNA and the color legend for each RNA biotype are the same as those shown in Figure 1. Asterisks denote significant differences between the ICH and IS transcriptomes for the particular transcript biotype. Bright orange arrow points to T-Cell Receptor genes.

Figure S9. Temporal profiles 10. A. Profile 10 in ICH only (there was no similar profile in IS). B. Profile 10 – in IS only (there was no similar profile in ICH).

Figure S10. Principal Components Analysis of 55 DET from T-cell receptor genes can distinguish patients with Intracerebral Hemorrhage (ICH) within 24 hours post-onset from: A. Ischemic Stroke (IS) patients within 24 hours post-onset. B. Control (CTRL) subjects. On Scan Date-, Age- and Sex-effect corrected data.

Figure S11. Principal Components Analysis of 107 DET involved in T-cell receptor function can distinguish patients with Intracerebral Hemorrhage (ICH) within 24 hours post-onset from: A. Ischemic Stroke (IS) patients within 24 hours post-onset. B. Control (CTRL)subjects. On Scan Date-, Age- and Sex-effect corrected data.

Figure S12. Time-dependent behavior of the 55 DET from T-cell receptor genes, which were differentially expressed in the acute phase (ICH TP1 vs TP0 (CTRL)).

Figure S13. Unsupervised Hierarchical Clustering of the 256 DET between ICH and IS. Subjects are in rows, DET are in columns. Red – high expression, green – low expression. Additional factors, such as biological sex, cause of IS and tPA application are displayed. DET – differentially expressed transcripts; CE IS (cardioembolic ischemic stroke); Lac IS (lacunar ischemic stroke); LV (large vessel ischemic stroke); ICH (intracerebral hemorrhage).

SUPPLEMENTARY TABLES

Supplementary Table 1. Subject Demographics and Clinical Characteristics.

Supplementary Table 2. Differentially Expressed Transcripts (DET). A. The 489 DET in ICH vs CTRL. B. The 63 DET in IS vs CTRL. C. The 396 DET in IS vs CTRL. D. The 256 DET in ICH vs IS.

Supplementary Table 3. Functional Analysis of the DET performed in Ingenuity Pathway Analysis (IPA). A. For the 489 DET from ICH vs CTRL. B. For the 396 DET from IS vs CTRL. C. For the 256 DET from ICH vs IS.

Supplementary Table 4. Differentially Expressed Transcripts (DET) in ICH and IS in the acute phase. A. The 2,667 DET in ICH_TP1 vs CTRL_TP0. B. The 311 DET in IS_TP1 vs CTRL_TP0.

Supplementary Table 5. Functional Analysis of the DET in the acute phase (0-24 hours, Time-point 1 (TP1) compared to Control (CTRL, TP0) performed in Ingenuity Pathway Analysis (IPA). A. For the 2,667 DET from ICH_TP1 vs CTRL_TP0. B. For the 311 DET from IS_TP1 vs CTRL_TP0.

Supplementary Table 6. Differentially Expressed Transcripts (DET) over time in ICH and IS – between each two time points.

Supplementary Table 7. Differentially Expressed Transcripts (DET) over time in ICH and IS – in each of the 10 profiles

Supplementary Table 8. Functional Analysis of the DET from each of the 10 profiles in ICH and in IS.

Supplementary Table 9. 107 DET of genes involved in T cell receptor function that are differentially expressed between ICH within 24 hours post-onset and CTRL.

Supplementary Table 10. Fisher's Exact p-values for the time-dependent bins. ICH - intracerebral hemorrhage; IS - ischemic stroke; CTRL - control. TP0 - control; TP1 - IS or ICH within 24 hours post event; TP2 - IS or ICH between 24 hours and 48 hours post event; TP3 - IS or ICH with >48 hours post event.

Supplementary Table 11. A. Complete Blood Count the day of the molecular microarray studies; Mann-Whitney test, 1,000 iterations. B. CIBERSORT deconvolution of whole blood microarray data; Mann-Whitney test, 1,000 iterations. ICH - intracerebral hemorrhage; IS - ischemic stroke; CTRL - control.

Supplementary Table 1. Subject Demographics and Clinical Characteristics.

	Intracerebral Hemorrhage	Ischemic Stroke	Vascular Risk Factor Controls	Total
Subjects, no.	33	33	33	99
Sex (Male, Female)	24, 9	24, 9	24, 9	72, 27
Age, years (Mean ± SD)	62 ± 14.3	64.8 ± 13.0	63.5 ± 13.1	63.4 ± 13.4
Race / Ethnicity (%)				
White	42.4	54.5	54.5	50.5
Black or African American	15.2	18.2	3.0	12.1
Asian	9.1	9.1	24.2	14.1
Native Hawaiian or Other Pacific Islander	0.0	3.0	0.0	1.0
Mixed Race	21.2	6.1	3.0	10.1
Latino	12.1	9.1	15.2	12.1
Time post-event, hours (Mean ± SD)	57.3 ± 30.6	48.5 ± 28.0	0	
Min	4.2	12	0	
Q1	30.6	23.1	0	
Q2	58.2	47	0	
Q3	83.6	68.2	0	
Max	124.3	110.4	0	
Hypertension, no.	23	25	23	71
Diabetes, no.	6	6	6	18
Hyperlipidemia, no.	6	12	12	30
Current Smoker, no.	7	9	5	21

SUPPLEMENTARY MATERIALS AND METHODS

1. Subjects

Ninety-nine male (M) and female (F) patients with intracerebral hemorrhage (ICH, n=33, 24M/9F), acute ischemic stroke (IS, n=33, 24M/9F), and vascular risk factor (VRF) - matched control subjects (CTRL, n=33, 24M/9F) were recruited between 2005 and 2013 from Universities of California at Davis and San Francisco, and at the University of Alberta, Canada. The protocol was approved by the UC Davis and UC San Francisco Institutional Review Boards and the University of Alberta Health Research Ethics Board and adheres to all federal and state regulations related to the protection of human research subjects, including The Common Rule, the principles of The Belmont Report, and Institutional Policies and procedures. Written informed consent was obtained from all participants or their proxy. ICH or IS diagnoses were made by board-certified neurologists based upon histories, exams, and computed tomographic (CT) brain scan and/or magnetic resonance imaging (MRI)[2]. The 33 IS subjects were represented by 11 cardioembolic, 11 large vessel, and 11 lacunar causes of IS. Five out of the 33 IS patients were treated with rtPA (recombinant tissue plasminogen activator) prior to the blood draw. CTRLs included subjects matched for VRFs and included stroke mimics, such as migraine and simple seizures. Samples were matched for age, race, sex, and VRFs, including hypertension, diabetes mellitus, hyperlipidemia, and smoking status. Exclusion criteria included previous stroke (for control subjects) and IS with hemorrhagic transformation.

2. Blood Collection, RNA Isolation

Blood collection and RNA isolation were performed as previously described[3]. There was a single blood draw per subject. For the IS and ICH subjects, time after symptom onset varied from 4.5 hours to 124.3 hours (Supplementary Table 1). For CTRL subjects, time was recorded as zero. Blood draw time was included as a covariate in the statistical analyses.

3. Array Design and Processing

The GeneChip® Human Transcriptome Array (HTA) 2.0 (Affymetrix, Santa Clara, CA) allows for deep examination into the coding and non-coding transcriptome and permits investigation of alternative splicing/transcript variants. It has an extensive coverage of the different layers of the transcriptome, with >285,000 full-length transcripts covered, of which >245,000 are coding transcripts, >40,000 - non-coding transcripts, and >339,000 - probe sets covering exon-exon junctions

(http://tools.thermofisher.com/content/sfs/brochures/hta_array_2_0_datasheet.pdf). Raw expression values (probe-level data) for each gene were saved in Affymetrix.CEL and Affymetrix.DAT files. Through Affymetrix Power Tools (APT), the .CEL transformer was used to apply GC Correction (GCCN) and Single Space Transformation (SST) to the HTA.CEL files. We conducted the GCCN-SST transformations through a command line using APT 1.18.0 in "batch" mode (runs the algorithm on groups of .CEL files rather than processing each individually) (https://www.affymetrix.com/support/developer/powertools/changelog/VIGNETTE-apt-cel-transformer-GCCN-SST.html). In order to investigate differential alternative splicing more directly, we performed transcript-level differential expression analysis using Partek Flow's workflow for HTA arrays. The intensity data is converted to simulate RNA-seq data by mapping within-exon and exon-exon junction probe sets onto the genome, and converting their intensities into "read" counts. For that, after performing the GCCN-SST transformations, the APT-preprocessed .CEL files were uploaded in the Partek Flow software (Partek Inc, St. Louis, MO). Probe sets were mapped to the Human Genome hg19, using STAR 2.4.1d aligner. Nominal read coverage depth was considered as 30 million and default mapping parameters were used. The mapping was followed by a quantification to a transcriptome (Ensembl75), RPKM normalization, and normalization offset of 1.0e-04 to account for the zero values. Low expression filter was applied (lowest maximum coverage=10). The E/M algorithm as implemented in Partek was used to assign "reads" (from the "mapping" of the HTA 2.0 probe sets) to known isoforms of a gene. Partek then uses a log-likelihood ratio test to identify DET across the groups.[4]

The quantified, normalized transcript-level expression values were analyzed using Mixed Regression Models in Partek Genomics Suite 6.6 (Partek Inc, St. Louis, MO). This approach allowed us to investigate alternative splicing/expression of transcript variants to detect **differentially expressed transcripts (DET)** between ICH and CTRL, IS and CTRL, and ICH and IS, as well as time-dependent expression following ICH and IS.

4. Analyses - Pairwise Group Comparisons

We used a Mixed Regression Model for transcript-level analyses: Diagnosis (ICH, IS, CTRL, fixed effect), Sex (categorical), Age (continuous), Time-since-event (continuous), Technical Variation (scan batch, random effect). REML variance estimate suitable for mixed models was used[5]. Transcripts with FDR-corrected p(Diagnosis)<0.2, which also had p<0.005 and |FC|>1.2 for the individual comparisons (ICH vs. CTRL, IS vs. CTRL, or ICH vs. IS), were

considered significant. The Fisher's least significant difference contrast method was used for the individual contrasts[6].

5. Analyses – Time-Dependent Changes in Transcript Expression

For time-dependent analyses (in a cross-sectional design), the VRF-matched controls (n=33) were considered as time-point 0 (TP0), and ICH and IS subjects were binned separately for time-since-event into additional three time-points – <24h (day 1, Time-Point 1 (TP1), 10 IS, 7 ICH), 24-48h (day 2, Time-Point 2 (TP2), 7 IS, 7 ICH) and >48h (≥day 3, Time-Point 3 (TP3), 16 IS, 19 ICH). Mixed Regression Model on transcript-level data was used: Diagnosis (IS, ICH, CTRL), Sex, Age, Time-Point (TP0, TP1, TP2, TP3), and an interaction of Diagnosis * Time-Point. Since a smaller number of subjects per group were analyzed in this analysis, significance was considered with $p<0.005$ and $|FC|>1.2$ on the individual contrasts from the interaction term.

6. *Self-Organizing Maps (SOM) Analyses*

For identifying similar profiles of dynamic changes following ICH and IS over the four time-points, we performed SOM[7] and Learning and Neighborhood functions, as implemented in Partek Genomics Suite. The initial learning rate was 0.1. The learning rate was updated with exponential decay. The neighborhood function type was Gaussian. SOM arranges data points in a map so that similar points are closer to each other on the map. The input for SOM was all transcripts that were differentially expressed between any two time-points ($p<0.005$, $FC>|1.2|$). Clustering was performed separately for the ICH and IS time-dependent transcripts. We performed 500,000 training iterations and set the Map Height to 4 and Map Width to 4, yielding 16 profiles for ICH and 16 - for IS. Each transcript-level's expression was normalized by standardizing it by shifting it to a mean of 0 and standard deviation (SD) of 1, and are presented summarized with ± 2SD in Figure 7 and Supplementary Figure 7.

7. Pathway Analysis. Prediction of Activation/Inhibition of Pathways. Identifying Upstream Transcriptional Regulators

Pathway enrichment analysis was performed in Ingenuity Pathway Analysis (IPA, Ingenuity Systems®) as previously described[8]. In short, Benjamini-Hochberg corrected $p<0.05$ was used for determining whether there were more genes per pathway that were differentially expressed between ICH and CTRL, IS and CTRL, or ICH and IS than would be expected by chance. For the time-dependent analysis (time profiles), Fisher's exact tests with nominal $p<0.05$ was used to ensure more detailed overview of the pathways involved.

IPA's Pathway Activity analysis allows to predict if a Signaling Canonical Pathway is increased (activated) or decreased (inhibited). For that, a Z-score algorithm is used to calculate a Z-score that mathematically compares our uploaded dataset of DET (by Gene Symbol) with the Canonical Pathway patterns. The latter ones are calculated taking into account the activation state of one or more key molecules when the pathway is activated, as well as the molecules' causal relationships with each other to generate an activity pattern for the molecules and the end-point functions in that pathway based on IPA-curated literature findings.

To identify potential upstream transcriptional regulators that may drive the observed transcriptome changes, we performed Upstream Transcriptional Regulators Analysis in IPA. It is based on prior knowledge of expected effects between transcriptional regulators and their target genes. As per IPA, for each potential transcriptional regulator, an overlap p-value (measures a significant overlap between our dataset genes and known targets regulated by a transcriptional regulator) and an activation z-score (infers likely activation/inhibition states of upstream regulators based on comparison with a model that assigns random regulation directions) are calculated for each upstream regulator.

8. Estimating the Fraction of Peripheral Blood Cell Types

To estimate the frequency of circulating immune cells for each subject, we utilized the Cibersort deconvolution method (https://cibersort.stanford.edu/)[9]. The method applies a machine learning algorithm (linear support vector regression (SVR)) to deconvolute unknown cell mixture based on an input matrix of reference gene expression signatures, collectively used to estimate the relative proportions of each cell type of interest[9]. The method is very robust compared to other methods with respect to noise and overfitting. We used the Cibersort's LM22 signature matrix, which contains 547 genes that distinguish 22 human hematopoietic cell phenotypes, including seven T cell types, naïve and memory B cells, plasma cells, NK cells, and myeloid subsets. Mann-Whitman test with 1,000 iterations was used to determine significance.

9. Biotype Comparisons

Fisher's Exact Test was used to calculate the difference in the numbers of DET in each biotype in the ICH vs CTRL analysis compared to the IS vs CTRL analysis. For each biotype category we input the "In-Biotype" and the "Not-in-Biotype" numbers for each biotype category to obtain the Fisher's Exact two-tailed p-value (GraphPad online calculator - https://www.graphpad.com/quickcalcs/contingency1/).

REFERENCES

1. Watkins NA, Gusnanto A, de Bono B, et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood* 2009; 113: e1-9. 2009/02/21. DOI: 10.1182/blood-2008-06-162958.
2. Yew KS and Cheng E. Acute stroke diagnosis. *Am Fam Physician* 2009; 80: 33-40. 2009/07/23.
3. Dykstra-Aiello C, Jickling GC, Ander BP, et al. Altered Expression of Long Noncoding RNAs in Blood After Ischemic Stroke and Proximity to Putative Stroke Risk Loci. *Stroke* 2016; 47: 2896-2903. 2016/11/12. DOI: 10.1161/STROKEAHA.116.013869.
4. Xing Y, Yu T, Wu YN, et al. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res* 2006; 34: 3150-3160. 2006/06/08. DOI: 10.1093/nar/gkl396.
5. Thompson WA, Jr. The problem of negative estimates of variance components. *Ann Math Stat* 33: 273–289.
6. Tamhane AC and Dunlop DD. *Statistics and Data Analysis: From Elementary to Intermediate*. Upper Saddle River, NJ: Prentice Hall, 2000.
7. Kohonen T. *Self-Organizing Maps*. 3rd edition ed.: Springer.
8. Stamova B, Green PG, Tian Y, et al. Correlations between gene expression and mercury levels in blood of boys with and without autism. *Neurotox Res* 2011; 19: 31-48. 2009/11/26. DOI: 10.1007/s12640-009-9137-7.
9. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015; 12: 453-457. 2015/03/31. DOI: 10.1038/nmeth.3337.