Am J Transl Res 2011;3(2):180-196
www.ajtr.org /ISSN:1943-8141/AJTR1101001

Original Article

ROC-supervised principal component analysis in connection with the diagnosis of diseases

Jason B. Nikas[1,2], Walter C. Low[1,3,4,5,6]

[1]Department of Neurosurgery, [2]Pharmaco-Neuro-Immunology Program, [3]Graduate Program in Neuroscience, [4]Department of Integrative Biology and Physiology, [5]Institute for Translational Neuroscience, [6]Center for Neuroengineering, Medical School, University of Minnesota, Minneapolis, MN, USA.

Received January 12, 2011; Accepted February 1, 2011; Epub February 3, 2011; Published February 15, 2011

Abstract: Principal component analysis (PCA) is a data analysis method that can deal with large volumes of data. Owing to the complexity and volume of the data generated by today's advanced technologies in genomics, proteomics, and metabolomics, PCA has become predominant in the medical sciences. Despite its popularity, PCA leaves much to be desired in terms of accuracy and may not be suitable for certain medical applications, such as diagnostics, where accuracy is paramount. In this study, we introduced a new PCA method, one that is carefully supervised by receiver operating characteristic (ROC) curve analysis. In order to assess its performance with respect to its ability to render an accurate differential diagnosis, and to compare its performance with that of standard PCA, we studied the striatal metabolomic profile of R6/2 Huntington disease (HD) transgenic mice, as well as that of wild type (WT) mice, using high field *in vivo* proton nuclear magnetic resonance (NMR) spectroscopy (9.4-Tesla). We tested both the standard PCA and our ROC-supervised PCA (using in each case both the covariance and the correlation matrix), 1) with the original R6/2 HD mice and WT mice, 2) with unknown mice, whose status had been determined via genotyping, and 3) with the ability to separate the original R6/2 mice into the two age subgroups (8 and 12 wks old). Only our ROC-supervised PCA (both with the covariance and the correlation matrix) passed all tests with a total accuracy of 100%; thus, providing evidence that it may be used for diagnostic purposes.

Keywords: Diagnostic methods, principal component analysis, receiver operating characteristic (ROC) curve analysis, metabolomics, nuclear magnetic resonance spectroscopy, huntington disease

Introduction

The concept of principal component analysis (PCA) was introduced by Pearson [1] and was developed by Hotelling [2-5]. Since then, PCA has been used in many research areas, including natural sciences, medical sciences, and behavioral and social sciences.

PCA is a multivariate data analysis/mining technique that seeks to transform, in a linear way, M correlated sets of P independent variables (IVs) into K uncorrelated sets of P IVs (where K<<M). The goal of PCA, in other words, is to reduce significantly the dimensionality of the original IVs (P) so that 1) the amount of the original variance accounted for by the number of the retained sets (K) of P IVs is maximized and 2) the K retained sets of P IVs are uncorrelated with each other. The fact that PCA is designed to replace a large number of sets of IVs with just a few (usually two or three) sets of those original IVs, with the condition that the few retained sets are not correlated, and also with the condition that those few retained sets capture the largest possible amount of the information (variance) contained in the original sets of IVs, has a significant and deterministic impact on both the applicability and performance of PCA.

Since the intended function of PCA is data dimensionality reduction, many have noted the advantages and disadvantages of PCA in that regard [3, 6-9]. Very little has been said, however, about PCA in connection with classification accuracy, a critical prerequisite for diagnostics. In this study, we investigated PCA specifically with respect to classification accuracy, as-

sessed its performance, and offered explanations about its evidenced weaknesses based on specific examples from our study (see section 3 of Supplementary Material). Moreover, and more importantly, in order to increase its classification accuracy and render it suitable for diagnostic applications, we introduced a new PCA method, one that is carefully supervised by receiver operating characteristic (ROC) curve analysis. Just as we did in the case of standard PCA, we used our nuclear magnetic resonance (NMR) spectroscopy study of Huntington disease (HD) in mice to assess the performance of the ROC-supervised PCA; and we compared the results with those of the standard PCA.

Brief Description of ROC-supervised PCA: 1) All of the variables of the original dataset are assessed in terms of their discriminating power between the target and the reference group (ROC AUC); 2) Those variables with an AUC > θ_1 (recommended θ_1 = 0.75) are used in the 1st PCA setting; 3) The classification results of the 1st PCA setting with respect to the original subjects according to the equation of the first principal component (PC_1) are recorded, and both the sum and the mean value of the squared residuals of every original subject as predicted by PC_1 (Q_1) are calculated; 4) Those variables with an AUC > θ_2 (recommended θ_2 = 0.80) are used in the 2nd PCA setting; 5) The classification results of the 2nd PCA setting with respect to the original subjects according to the equation of the first principal component (PC_1) are recorded, and both the sum and the mean value of the squared residuals Q_1 are calculated; 6) The previous two steps are repeated k times with increasing AUC values until the k^{th} PCA setting, wherein only those original variables with an AUC > θ_k are used, yields a) the most accurate classification results with respect to the original subjects and b) the smallest mean value and sum value of all Q_1 squared residuals. This k_{th} PCA setting constitutes the diagnostic model; 7) The diagnostic model is tested with unknown subjects.

Materials & methods

R6/2 transgenic mice

Animal experiments described in this study were performed in accordance to the procedures approved by the University of Minnesota Institutional Animal Care and Use Committee. The R6/2 mice were originally purchased from the Jackson Laboratories (Bar Harbor, ME, USA) and bred by crossing transgenic males and wild type (WT) females at 5 weeks of age. Offspring were genotyped according to established procedures [10] and the Jackson Laboratory.

Animal preparation

In preparation for *in vivo* ^{1}H NMR (proton nuclear magnetic resonance) scanning, all animals were anesthetized and maintained thus throughout the duration of the scanning procedure. A gas mixture (O_2 : N_2O = 1:1) containing 1.25–2.0% of isoflurane was used for anesthesia and flowed throughout the cylindrical chamber wherein the spontaneously breathing animals were placed. The chamber temperature was maintained at 30^0 C by the circulation of warm water on the outside surface of the chamber. The ^{1}H NMR scanning for each animal required approximately 1hr.

In Vivo ^{1}H NMR spectroscopy

^{1}H NMR scans were conducted with a 9.4 T/31 cm magnet (Magnex Scientific, Abingdon, UK). The magnet was equipped with an 11 cm gradient coil insert (300 mT/m, 500 ls) and strong custom-designed second order shim coils (Magnex Scientific, Abingdon, UK) [11]. The volume of interest (VOI) was selected based on multi-slice RARE images. The VOI was centered in the left striatum at the level of the anterior commissure. The size of the VOI, which varied from7–12 μL, was adjusted to fit the anatomical structure of the left striatum, as well as to exclude the lateral ventricle and, thus, to minimize partial volume effects (inclusion of a tissue other than the target tissue). The striatum was selected as the area of interest because it consists to a large extent of the medium spiny projection neurons, which are GABA-ergic, and which, more importantly, constitute the initial and preferential target of HD. It is in the medium spiny projection neurons of the striatum where HD first manifests itself. At the end stage, following extensive neuronal cell loss in the striatum, the disease evinces itself in other brain areas, such as the cerebral cortex, globus pallidus, substantia nigra, thalamus, cerebellum, nucleus accumbens, and white matter [12].

Thirty mice (17 WT and 13 R6/2) were scanned according to the aforementioned procedure. Of the 17 WT mice, 8 were 8 wks old and 9 were

12 wks old; whereas of the 13 R6/2 mice, 7 were 8 wks old and 6 were 12 wks old. Those 30 mice were used in the development of both the standard and the ROC-supervised PCA diagnostic biomarker models (DBMs). In addition, 31 unknown mice (11 R6/2 and 20 WT) were also scanned according to the aforementioned procedure and were used to test and validate all PCA DBMs. All of the 31 unknown mice were extraneous to the development of the PCA DBMs, and their status had been ascertained via genotyping.

Spectral analysis resulted in the identification and individual quantification of 15 metabolites. By combining the obtained individual absolute concentrations of creatine (Cr) and phosphocreatine (PCr), we created the Cr+PCr and PCr/Cr metabolites (variables) in order to obtain information about the total striatal creatine (free and phosphorylated), as well as about the ratio of those two metabolites. In the case of glycerophosphorylcholine (GPC) and phosphorylcholine (PC), we were not able to separate those two and obtain individual concentrations. We were able, however, to obtain the absolute concentration of the sum of GPC and PC, which represents the total striatal phosphorylated choline. All of the 15 striatal metabolites we were able to identify and quantify individually as a result of the high magnetic field spectrometer we used (9.4 Tesla), as well as the two metabolites (variables) we created, are shown in **Table 1**.

Since both of our animal groups (WT & R6/2) comprised two age subgroups (8-wk old & 12-wk old mice), the time dependent variable was collapsed, so the developed models for diagnostic biomarkers (DBMs) would be applicable from 8-12 weeks of age – a most important time period in the progression of the disease in R6/2 mice, as well as a significant portion of the observed lifespan of the R6/2 mice. The development of all diagnostic biomarker models (DBMs), therefore, was based on the data of the aforementioned 13 R6/2 mice [seven at 8 wks of age & six at 12 wks of age] and 17 WT mice [eight at 8 wks of age & nine at 12 wks of age]. For more details on animal methods, as well as on spectra obtainment and processing, please see our previous study [13].

Statistical software

For our study, we used the statistical software by NCSS 2007, Kaysville, Utah, USA.

Table 1. Names & abbreviations of all metabolites detected and measured in the study

No.	Metabolite Symbol	Metabolite Name
1	Cr	creatine
2	PCr	phosphocreatine
3	Cr+PCr	creatine + phosphocreatine
4	PCr/Cr	phosphocreatine / creatine
5	GABA	γ-aminobutyric acid
6	Glc	glucose
7	Gln	glutamine
8	Glu	glutamate
9	GSH	glutathione
10	GPC+PC	glycerophosphorylcholine + phosphorylcholine
11	Lac	lactate
12	MM	macromolecules
13	mIns	*myo*-Inositol
14	NAA	N-acetylaspartate
15	NAAG	N-acetylaspartylglutamate
16	PE	phosphorylethanolamine
17	Tau	Taurine

Computer programs

Computer programs were written using MATLAB R2009b by The MathWorks, Inc., Natick, MA, USA.

Diagnostic biomarker models

General Description: We used the data (concentrations of 17 metabolites) of our 30 original mice to develop diagnostic biomarker models (DBMs) for both the standard and the ROC supervised PCA methods. The DBMs comprised computer programs, which, based on the equation of the first principal component (PC_1) (1.1) of the respective PCA method, could render a differential diagnosis of an unknown mouse (WT or R6/2). More specifically, the equation of the first principal component is given by

$$PC_{1N} = w_{11}X_{1N} + w_{12}X_{2N} + \ldots + w_{1P}X_{PN} \qquad (1.1)$$

X_{1N}, X_{2N}, ..., X_{PN} are the P variables (in our case, P=17 metabolite concentrations) of subject N; w_{11}, w_{12}, ... w_{1P} are the weights of the P variables with respect to PC_1, which can be calculated from the eigenvector of PC_1; and PC_{1N} is the score of subject N with respect to the first principal component (PC_1). The first principal component (PC_1) is the most important of all

principal components for the following two reasons: 1) it contains most of the information (variance) of the original variables and 2) it has the highest potential in terms of classification accuracy with respect to the target and the reference group (see results in Tables S1-S10 in the Supplementary Material). Therefore, we can use equation (1.1) to make a diagnosis of an unknown mouse by calculating its score with respect to the first principal component. Based on whether the score is positive or negative, the unknown mouse can be diagnosed as either WT or R6/2 respectively. More details on (1.1) and other PCA equations, as well as the basic theory of PCA, can be found in section 1 of Supplementary Material.

We subjected both PCA DBMs (standard and ROC-supervised) to the following three tests:

Test 1: Identification of our original 30 mice, which were intrinsic to the development of all DBMs. This is a necessary first test in that a DBM has to demonstrate that it has the prerequisite discriminating accuracy to classify correctly the original 30 mice, which were used in the development of that DBM. It is by no means a foregone conclusion that a DBM can pass this test with 100% accuracy.

Test 2: Identification of 31 unknown mice, which were extraneous to the development of all DBMs. This is the validation test, and as such, it is by far the most important test. A DBM is asked to identify/diagnose 31 unknown mice. These 31 mice were new and different from the 30 original mice used in the development of that DBM. The status of these 31 unknown mice had been determined by genotyping, which is the gold standard in HD.

Test3: Identification of our 13 original R6/2 mice into their two age groups: 8 wk-old and 12 wk-old. Seven of those R6/2 mice were scanned at the age of 8 weeks and six of them were scanned at the age of 12 weeks. This is a test designed to assess the sensitivity of a DBM with respect to the progression of the disease. Those R6/2 mice that were scanned at the age of 12 weeks were more impaired than those R6/2 mice that were scanned when they were 8 weeks old. A DBM should have the required sensitivity to discriminate between those two groups of R6/2 mice.

For both the standard and the ROC-supervised PCA in connection with the first test, we entered our data (subjects) in the following order: rows #1-17 were the WT mice and rows #18-30 were the R6/2 mice. For both the standard and the ROC-supervised PCA in connection with the third test, we entered our data (subjects) in the following order: rows #1-7 were the 8-wk old R6/2, whereas rows #8-13 were the 12-wk old R6/2 mice.

PCA with Covariance Matrix: For both the standard and the ROC-supervised PCA with the covariance matrix, we chose the following settings: Matrix Type: We chose the Covariance Matrix; Factor Selection – Method: We chose Percent of Eigenvalues; Factor Selection – Value: We selected 100; Factor Rotation: We chose none.

PCA with Correlation Matrix: Except for the Matrix Type, for the correlation matrix PCAs (both the standard and the ROC-supervised one), we chose the same settings as those for the PCAs with the covariance matrix (listed immediately above). In section 2 of the Supplementary Material, there is an account of the differences between PCA with covariance matrix and PCA with correlation matrix.

ROC curve analysis

ROC curve analysis is a theory of probabilities. It studies two probabilities, namely, sensitivity and (1-specificity), in order to determine a third probability, namely, the area under the curve (AUC). The ROC AUC probability is basically an assessment of the discriminating power of a given variable with respect to the two groups involved. If the AUC of a given variable is equal to 1.00, then according to that variable, the two groups involved can be separated with 100% accuracy. A variable with perfect discrimination between the two groups has an AUC = 1.00, whereas a variable with the poorest discrimination between the two groups has an AUC = 0.50 (chance probability). For a more detailed account on the properties, methodology, and applications of ROC curve analysis, please refer to our previous study [14].

Since ROC curve analysis allows us to assess our variables in terms of discriminating power with respect to our two groups (WT vs. R6/2), we used the results of ROC curve analysis (**Table 2**) not only to supervise PCA but also to determine the best possible setting of the ROC-supervised PCA. To be more specific, first we

Table 2. Rank of all metabolites based on their discriminating power (AUC) from ROC curve analysis

Time: 8-12 wks	ROC Curve Analysis	
Metabolite	AUC	AUC Rank
Cr+PCr	1.00000	1
Gln	0.98897	2
Cr	0.98832	3
NAA	0.98198	4
GSH	0.94052	5
GPC+PC	0.90301	6
mIns	0.89978	7
PCr	0.87023	8
PE	0.83667	9
Tau	0.72888	10
NAAG	0.69632	11
Glc	0.58495	12
Glu	0.58179	13
PCr/Cr	0.53852	14
GABA	0.52209	15
Lac	0.52187	16
MM	0.50067	17

entered only those IVs (metabolite concentrations) that had an AUC > 0.70 (70%), then only those that had an AUC > 0.80, then only those with an AUC > 0.90, and finally only those with an AUC > 0.95. In order to assess the different settings of the ROC-supervised PCA, we used the following criteria: 1) classification results and 2) the residuals (both the sum of the Q_1 values of all subjects and the mean Q_1 value of all subjects of a particular setting). Q_P is the sum of squared residuals when a subject is predicted using the first P principal components [4]. Since we are interested in the first principal component, Q_1 is the residual of our interest. The smaller the sum of all the residuals of all subjects (sum of all Q_1 values of all subjects) and the smaller the mean value of the residuals of all subjects (mean value of the Q_1 values of all subjects), the better the setting.

Results

Standard PCA with covariance matrix

Test 1: Identification of the original 30 mice (WT vs. R6/2): We ran the standard PCA with the covariance matrix and unsupervised, i.e. with all of our 17 IVs (metabolites). Rows # 1-17 were the WT mice, and rows #18-30 were the R6/2 mice. Table S1 in the Supplementary Material shows the scores of the 30 original mice with respect to the first test according to the first six principal components (factors) (PC_1 – PC_6). None of the 17 principal components correctly identified all of the 30 mice. As one can see from Table S1, the first principal component (PC_1) misidentified 5 mice (#19-22 & #24), which are R6/2, and which should have negative factor scores. The results, therefore, according to PC_1 are: 17/17 WT mice (100% correct) & 8/13 R6/2 mice (61.54% correct) → with a total accuracy of 25/30 original mice (83.33% correct). In this case, sensitivity = 0.615 and (1-specificity) = 0.

The positive Likelihood Ratio [(+)LR] is: (+)LR = (sensitivity)/(1-specificity) = 0.615/0 → ∞ The negative Likelihood Ratio [(-)LR] is: (-)LR = (1-sensitivity)/(specificity) = 0.385/1 = 0.385.

As can be seen in **Figure 1**, there is no separation between the WT mice (#1-17) and the R6/2 mice (#18-30) either with respect to PC_1 or PC_2. The general results of all PCA runs, including those of this run, appear in **Table 3**. The second principal component (PC_2) misidentified 6 mice: #26 & #28-30, which are R6/2, and which should have negative factor scores, as well as

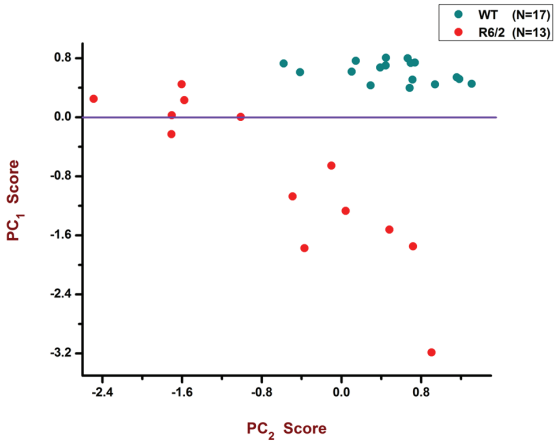

Figure 1. Standard PCA (covariance matrix) – Test 1. Scores of the 30 original mice according to the first principal component (PC_1 Score) plotted against the scores of the same mice according to the second principal component (PC_2 Score). The PCA was run unsupervised (all 17 IVs were used) using the covariance matrix. As can be seen, there is no separation between the two groups [WT (#1-17) & R6/2 (#18-30)] either with respect to the first principal component or with respect to the second one.

Table 3. General results of all PCA runs with respect to our three tests

PCA RESULTS				
	PCA COVARIANCE MATRIX		PCA CORRELATION MATRIX	
	Standard (17 Variables)	ROC-Supervised (4 Variables)	Standard (17 Variables)	ROC-Supervised (4 Variables)
Test 1: ID of original 30 mice	% Correct		% Correct	
17 WT	17/17 (100%)	17/17 (100%)	17/17 (100%)	17/17 (100%)
13 R6/2	8/13 (61.54%)	13/13 (100%)	13/13 (100%)	13/13 (100%)
Total	25/30 (83.33%)	30/30 (100%)	30/30 (100%)	30/30 (100%)
(+) Likelihood Ratio	$0.615/0 \rightarrow \infty$	$1/0 \rightarrow \infty$	$1/0 \rightarrow \infty$	$1/0 \rightarrow \infty$
(-) Likelihood Ratio	0.385	0/1=0	0/1=0	0/1=0
TEST 2: ID of 31 unknown mice				
20 WT	20/20 (100%)	20/20 (100%)	20/20 (100%)	20/20 (100%)
11 R6/2	7/11 (63.64%)	11/11 (100%)	8/11 (72.73%)	11/11 (100%)
Total	27/31 (87.10%)	31/31 (100%)	28/31 (90.32%)	31/31 (100%)
(+) Likelihood Ratio	$0.636/0 \rightarrow \infty$	$1/0 \rightarrow \infty$	$0.727/0 \rightarrow \infty$	$1/0 \rightarrow \infty$
(-) Likelihood Ratio	0.364	0/1=0	0.273	0/1=0
TEST 3: ID of original 13 R6/2 mice		(R6/2)-ROC-Supervised (2 Variables)		(R6/2)-ROC-Supervised (2 Variables)
7 R6/2 8 wks old	6/7 (85.71%)	7/7 (100%)	7/7 (100%)	7/7 (100%)
6 R6/2 12 wks old	6/6 (100%)	6/6 (100%)	6/6 (100%)	6/6 (100%)
Total	12/13 (92.31%)	13/13 (100%)	13/13 (100%)	13/13 (100%)
(+) Likelihood Ratio	6.998	$1/0 \rightarrow \infty$	$1/0 \rightarrow \infty$	$1/0 \rightarrow \infty$
(-) Likelihood Ratio	0/0.857=0	0/1=0	0/1=0	0/1=0

mice #2-3, which are WT, and which should have positive factor scores (Table S1). Not surprisingly, the rest of the factors (3-17), which collectively account for only ~ 20% of the original variance (Table S2 in the Supplementary Material), did not show any meaningful results with respect to the identification of the 30 mice. The individual significance of the 17 IVs for each of the first five principal components can be seen in **Table 4**. More specifically, the eigenvectors of the first 5 principal components (factors) of standard PCA (using the covariance matrix) are shown. Since the magnitude of the absolute value of the weights of the variables within each eigenvector is directly proportional to the significance of the variables for each principal component, one can see the magnitude of significance of each variable for each of the first five principal components. Focusing on PC_1 (Factor 1), one can see that Tau has by far the greatest weight (0.6756), and it is, therefore, the most significant variable for PC_1, which, in turn, is the most significant of all principal components since it alone accounts for 57.51% of the original variance (Table S2 in the Supplementary Material). That means that the equation of PC_1 has been heavily influenced by Tau. As can be seen in **Table 2**, Tau, according to ROC curve analysis, has an AUC = 0.7289, which means that in this case, Tau as a biomarker cannot be used for diagnostic purposes. Cr+PCr, on the other hand, is the perfect biomarker (AUC = 1.0000) (**Table 2**), and it is upon this variable (Cr + PCr) that the equation of PC_1 should have been predominantly based. In Section 3 in the Supplementary Material, there is a more detailed account and discussion of the two aforementioned metabolites in connection with the basic principle of operation of PCA.

It is elucidating to observe that the order of significance of the 17 IVs according to the eigenvector of PC_1 is as follows: 1) Tau, 2) Gln, 3) Cr+PCr, 4) Lac, 5) GPC+PC, 6) PCr, 7) NAA, 8)

Table 4. The eigenvectors of the first 5 principal components (factors) of standard PCA using the covariance matrix

Eigenvectors Variables	Factors				
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Cr	-0.140814	-0.404184	-0.252721	0.084649	0.003137
Gln	-0.421211	-0.291823	-0.023370	0.211430	0.324034
NAA	0.186569	0.208104	0.245636	0.080688	0.375307
Cr+PCr	-0.345292	-0.395072	-0.198341	0.044874	0.025746
PCr	-0.204489	0.009019	0.054307	-0.039749	0.022740
Glc	0.073144	-0.274866	0.010358	-0.918090	0.057282
Glu	-0.158447	0.169657	0.142106	-0.020456	0.629347
GSH	-0.051790	-0.059521	-0.063733	-0.032056	0.017220
mIns	-0.130164	-0.219076	0.048181	-0.045341	0.260795
Lac	0.213520	-0.517945	0.777880	0.163626	-0.129195
PE	0.088775	0.067588	-0.021296	-0.093113	0.246811
Tau	-0.675590	0.347693	0.427139	-0.205028	-0.274482
GPC+PC	-0.213405	-0.010141	-0.018493	0.085112	-0.177403
MM	0.003147	-0.011569	0.038310	-0.030734	-0.020392
GABA	-0.021660	0.014303	0.123482	-0.061108	0.316906
NAAG	-0.024488	-0.020983	0.015875	-0.043573	-0.004801
PCr/Cr	-0.017543	0.039048	0.031111	-0.018325	-0.000205

The magnitude of the absolute value of the weights of the variables within each eigenvector (column) is directly proportional to the significance of the variables for each factor. As can be seen from the absolute value of the weights of Factor 1 (PC_1), Tau has by far the greatest weight (0.675590), and it is, therefore, the most significant variable for Factor 1, which is the most significant of all factors as by itself it accounts for 57.51% of the original variance. That means that the equation of Factor 1 has been heavily influenced by Tau. The rest most significant variables for Factor 1 are Gln, Cr+PCr, Lac, GPC+PC, PCr, etc in a descending order of significance after Tau. All 17 IVs were used in this run.

Glu, 9) Cr, 10) mIns, 11) PE, 12) Glc, 13) GSH, 14) NAAG, 15) GABA, 16) PCr/Cr, and 17) MM. This order of significance of the 17 IVs is markedly different from that yielded by ROC curve analysis (**Table 2**). Besides the problem with the ranking of Tau, Lac, which has an AUC = 0.5219 (**Table 2**), which in essence means that the diagnostic (discriminating) power of Lac is at the chance level (AUC = 0.50), is ranked by standard PCA (covariance matrix) in the top four most significant metabolites. On the other hand, Cr, which has an AUC = 0.9883, which is considered excellent (> 0.95), is ranked by standard PCA as number 9 (out of 17).

Table S2 in the Supplementary Material shows the eigenvalues of the eigenvectors of all 17 principal components (factors) of standard PCA using the covariance matrix. As can be seen, the first principal component (PC_1) accounts for 57.51% of the original variance of the data; the second one (PC_2) accounts for 22.29%; the third one (PC_3) for 8.35%, etc. It is worth noting that the first four principal components collectively account for 91.53% of the original variance. Another observation that is worth mentioning is that since an eigenvalue represents the variance of a principal component, it can be seen that PC_1 has the largest variance (9.6213) of all principal components.

Test 2: Identification of the 31 unknown mice (WT vs. R6/2) – Validation Test: We subjected the standard PCA (covariance matrix) to the second test (identification of 31 unknown mice). Based on the equation of the first principal component we derived from the standard PCA (covariance matrix) in the previous Section (3.1.1), we wrote a computer program that, following the input of the 17 metabolite concentrations of an unknown mouse, would render a differential diagnosis as to whether that unknown mouse was a WT or an R6/2 mouse. As we mentioned previously, we had 31 unknown mice (11 R6/2 and 20 WT), which were extraneous to all of the DBMs, and the status of which had been determined via genotyping. Standard PCA with the covariance matrix correctly determined the status of 27/31 unknown mice [20/20 WT mice (100% correct) and 7/11 R6/2 mice (63.64% correct), with a total accuracy of 27/31 unknown mice (87.10% correct)]. There-

Table 5. Detailed results of all PCA runs with respect to the second test, i.e. the identification of the 31 unknown mice. Mice in rows #1-20 are WT, whereas mice in rows #21-31 are R6/2

PCA Results for Test 2				
Unknown Subject	First Principal Component (PC_1) Score			
	PCA Covariance Matrix		PCA Correlation Matrix	
	Standard	ROC-Supervised	Standard	ROC-Supervised
1	0.13167	0.61446	0.40658	0.78912
2	0.59647	0.32887	0.62741	0.34816
3	0.61650	0.58706	0.67466	0.68506
4	0.27976	0.64139	0.48024	0.84950
5	0.26657	0.93238	0.56482	1.20830
6	0.64921	0.99697	0.72184	1.21220
7	0.58748	0.82755	0.68411	0.97334
8	0.45766	0.89058	0.67364	1.03990
9	0.43063	0.94261	0.71080	1.09190
10	0.76189	1.08260	1.08250	1.24480
11	0.40067	1.03740	0.74022	1.21150
12	0.71320	1.24290	1.08390	1.49750
13	0.44182	1.16940	0.77265	1.37920
14	0.40866	1.15300	0.79191	1.28990
15	0.48321	0.77448	0.69302	0.97603
16	0.28809	0.76857	0.55896	0.92809
17	0.81518	1.01340	1.02390	1.19510
18	0.25073	0.54776	0.59821	0.55057
19	0.23476	1.01100	0.62662	1.23000
20	0.73047	0.95328	0.97916	1.13380
21	-0.89293	-1.15940	-0.78696	-1.43120
22	-0.91048	-1.02050	-1.14860	-1.41420
23	-1.78630	-1.77620	-1.76590	-2.02740
24	-1.03060	-1.68840	-1.29770	-2.03200
25	-1.24270	-1.79300	-1.43730	-2.14990
26	0.33614	-0.94422	-0.04884	-1.33740
27	-1.11610	-1.51380	-1.50130	-1.82760
28	0.61648	-0.01065	0.36699	-0.10788
29	0.11753	-0.30609	0.03627	-0.47616
30	0.88638	-0.31285	0.41756	-0.51304
31	-0.13088	-0.27708	-0.27506	-0.37472

Only the ROC-supervised PCA (both with the covariance and the correlation matrix) diagnosed/identified correctly all of the 31 unknown mice. All of the WT mice (rows # 1-20) have positive PC1 scores, whereas all of the R6/2 mice (rows # 21-31) have negative PC1 scores.

fore, for the second test, the standard PCA (covariance matrix) exhibited a sensitivity = 0.636 and a (1-specificity) = 0 [(+)LR = 0.636/0 $\rightarrow \infty$ and (-)LR = 0.364]. Detailed results of all PCA runs (standard and ROC-supervised) with respect to the second test are shown in **Table 5**. The general results of all PCA runs, including those of this run, appear in **Table 3**.

Test 3: Identification of the 13 original R6/2 mice (8-wk old vs. 12-wk old): Next, we subjected the standard PCA (covariance matrix) to our third test. More specifically, we wanted to know whether standard PCA (covariance matrix) was sensitive enough to detect the metabolomic differences caused by the progression of Huntington disease (HD) between our two R6/2 subgroups, i.e. between the 8-wk old R6/2 and the 12-wk old R6/2 mice. Physiologically, we know that the progression of HD will effect alterations in the metabolite concentrations of the cells in the striatum area of the

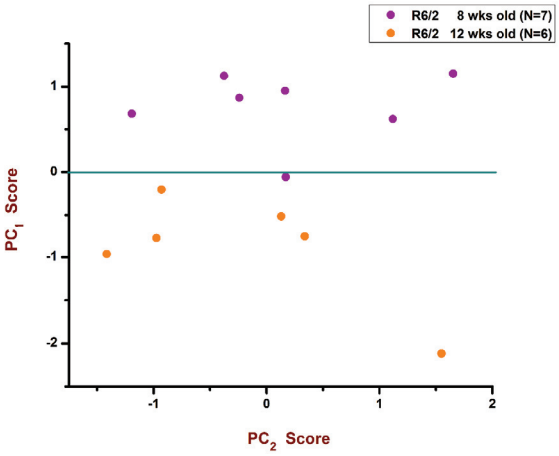

Figure 2. Standard PCA (covariance matrix) – Test 3. Scores of the 13 original R6/2 mice according to the first principal component (PC_1 Score) plotted against the scores of the same mice according to the second principal component (PC_2 Score). The PCA was run unsupervised (all 17 IVs were used) using the covariance matrix. As can be seen, there is no accurate separation between the two groups [8 wk-old R6/2 (#1-7) and 12 wk-old R6/2 (#8-13)] either with respect to the first principal component or with respect to the second one.

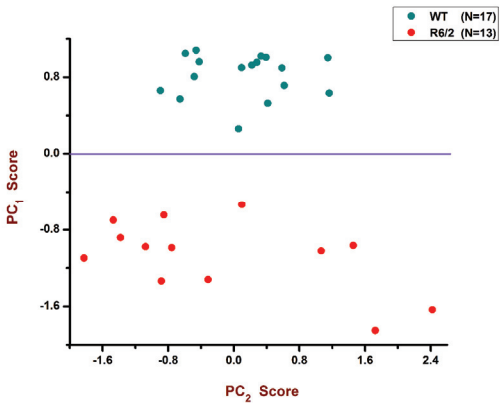

Figure 3. ROC-supervised PCA (covariance matrix) – Test 1. Scores of the 30 original mice according to the first principal component (PC_1 Score) plotted against the scores of the same mice according to the second principal component (PC_2 Score). The PCA was run using the covariance matrix and was supervised by the ROC curve analysis [the top four most significant IVs (Cr+PCr, Gln, Cr, and NAA) (AUC > 95%) as determined by the ROC curve analysis were used]. As can be seen, there is a separation between the two groups [WT (#1-17) & R6/2 (#18-30)] only with respect to the first principal component: all of the WT mice have positive scores, whereas all of the R6/2 mice have negative scores.

brain. A mathematical model, therefore, should be sensitive enough to detect those alterations in the time span of four weeks. We entered all of the 17 IVs and our 13 original R6/2 mice [(7) 8-wk old & (6) 12-wk old] in the following manner: rows #1-7 the seven 8-wk old ones and rows #8-13 the six 12-wk old ones. Standard PCA (covariance matrix) correctly identified and classified 12/13 of our original R6/2 mice into their respective two subgroups [6/7 8 wk-old R6/2 mice (85.71% correct) & 6/6 12 wk-old R6/2 mice (100% correct) → with a total accuracy of 12/13 original R6/2 mice (92.31% correct)]. In this case, the sensitivity = 1 and the (1-specificity) = 0.143 [(+)LR = 1/0.143 = 6.998 and (-)LR = 0/0.857 = 0]. As can be seen in **Figure 2**, there is no accurate separation of our two R6/2 groups either with respect to PC_1 or PC_2. The general results of all PCA runs, including those of this run, appear in **Table 3**.

ROC-supervised PCA with covariance matrix

Test 1: Identification of the original 30 mice (WT vs. R6/2): Our goal was to find the best ROC-supervised PCA setting for the 30 original mice and use that setting to develop a ROC-supervised DBM. Using the covariance matrix, we ran the PCA with the top 10 IVs (AUC > 70%), top 9 IVs (AUC > 80%), top 7 IVs (AUC ≥ 90%), and top 4 IVs (AUC > 95%) according to ROC curve analysis (**Table 2**). Of those runs, the last three correctly identified all of the 30 original mice [17/17 WT mice (100% correct) & 13/13 R6/2 mice (100% correct) à with a total accuracy of 30/30 original mice (100% correct)]. The run with the top 9 IVs (AUC > 80%) yielded a sum of all Q_1 residuals equal to 111.82, and a mean Q_1 residual value of 3.73. The corresponding values of the run with the top 7 IVs (AUC ≥ 90%) were: 75.17 and 2.51. The run with the top 4 IVs (AUC > 95%) yielded the following values respectively: 23.11 and 0.77. Clearly, the run with the top 4 IVs (AUC > 95%) was the best ROC-supervised PCA setting (covariance matrix) for the 30 original mice, and it was upon the equation of the first principal component of this setting that the ROC-supervised PCA DBM (covariance matrix) was based. As was mentioned, the ROC-supervised

PCA DBM (covariance matrix) correctly identified all of our 30 original mice [17/17 WT mice (100% correct) & 13/13 R6/2 mice (100% correct) → with a total accuracy of 30/30 original mice (100% correct)] [sensitivity = 1; (1-specificity) = 0; (+)LR = 1/0 → ∞; (-)LR = 0/1 = 0]. **Figure 3** depicts those results. As can be seen, our two groups were successfully separated (correctly identified) by the first principal component: all of the WT mice have positive scores, whereas all of the R6/2 mice have negative scores. Once again, **Table 3** depicts the general results of all of the PCA runs.

Test 2: Identification of the 31 unknown mice (WT vs. R6/2) – Validation Test: We subjected the ROC-supervised PCA DBM (covariance matrix) [using only the top 4 IVs (AUC > 98%) according to ROC curve analysis] to the second test. It correctly determined the status of all of the 31 unknown mice [20/20 WT mice (100% correct) and 11/11 R6/2 mice (100% correct), with a total accuracy of 31/31 unknown mice (100% correct)] [sensitivity = 1; (1-specificity) = 0; (+)LR = 1/0 → ∞; (-)LR = 0/1 = 0]. Those results in detail, along with the results of all PCA runs with respect to the second test, are shown in **Table 5**. The general results of all PCA runs, including those of this run, appear in **Table 3**.

Test 3: Identification of the 13 original R6/2 mice (8-wk old vs. 12-wk old): Subjecting the best ROC-supervised PCA setting to the third test was the next task. The third test concerns itself exclusively with the R6/2 mice; more specifically, it assesses the ability of a given model to discriminate between the two R6/2 groups: the 8-wk old vs. the 12-wk old. The ROC curve analysis with which we supervised PCA in the first and second test, and the results of which appear in **Table 2**, was designed to assess the ability of all 17 IVs to discriminate between the WT and the R6/2 mice. Clearly, as far as the third test was concerned, we had to perform another ROC curve analysis, one that would deal exclusively with the 13 original R6/2 mice, and one that would assess all of the 17 IVs in terms of their ability to discriminate between the 8-wk old and the 12-wk old R6/2 mice. The top 5 most significant IVs (metabolites) in the discrimination between the two R6/2 groups according to their AUC value as determined by the R6/2 ROC curve analysis are: 1) TTau (AUC = 0.9752) [Transformed Tau in order to meet normality criteria], 2) GPC+PC (AUC = 0.9517), 3) Glu (AUC = 0.9460), 4) Lac (AUC = 0.9446), and 5) Gln (AUC = 0.9432). The best R6/2 ROC-supervised PCA setting for the 13 original R6/2 mice both in terms of classification accuracy and residuals was the one that employed only the top two most significant IVs (AUC > 95%), i.e. TTau and GPC+PC; and it is this setting that we used for the third test. This R6/2-ROC-supervised PCA (covariance matrix) correctly identified and classified all of our original R6/2 mice into their respective two subgroups [7/7 8 wk-old R6/2 mice (100% correct) & 6/6 12 wk-old R6/2 mice (100% correct) → with a total accuracy of 13/13 original R6/2 mice (100% correct)] [sensitivity = 1; (1-specificity) = 0; (+) LR = 1/0 → ∞; (-)LR = 0/1 = 0]. **Figure 4** depicts those results. As can be seen, our two R6/2 groups were successfully separated (correctly identified) by the first principal component: all of the 8 wk-old R6/2 mice have positive scores, whereas all of the 12 wk-old R6/2 mice have negative scores. Those results are

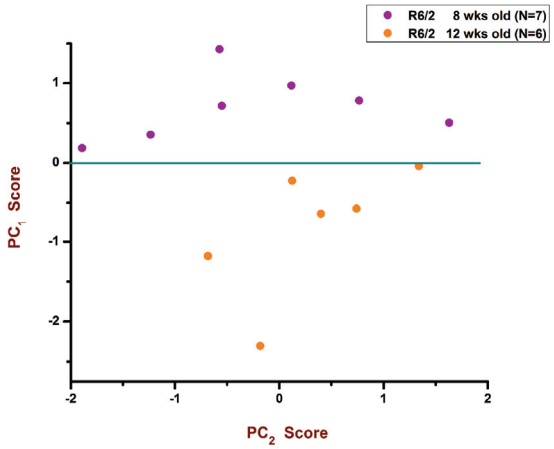

Figure 4. R6/2-ROC-supervised PCA (covariance matrix) – Test 3. Scores of the 13 original R6/2 mice according to the first principal component (PC_1 Score) plotted against the scores of the same mice according to the second principal component (PC_2 Score). The PCA was run using the covariance matrix and was supervised by the R6/2-ROC curve analysis [the top two most significant IVs (TTau and GPC+PC) (AUC > 95%) as determined by the R6/2-ROC curve analysis were used]. As can be seen, there is a separation between the two groups [8 wk-old R6/2 (#1-7) and 12 wk-old R6/2 (#8-13)] only with respect to the first principal component: all of the 8 wk-old R6/2 mice have positive scores, whereas all of the 12 wk-old R6/2 mice have negative scores.

also shown, along with the general results of all PCA runs, in **Table 3**.

Standard PCA with correlation matrix

Test 1: Identification of the original 30 mice (WT vs. R6/2): We next ran standard PCA (all 17 IVs) with the correlation matrix. As can be seen from Table S6 in the Supplementary Material, this PCA run was more successful than the standard PCA with the covariance matrix. More specifically, the first principal component (PC_1) correctly identified all of our 30 original mice: all WT mice have positive PC_1 scores, whereas all R6/2 mice have negative PC_1 scores [17/17 WT mice (100% correct) & 13/13 R6/2 mice (100% correct) $\rightarrow$ with a total accuracy of 30/30 original mice (100% correct)] [sensitivity = 1; (1-specificity) = 0; (+)LR = 1/0 $\rightarrow \infty$; (-)LR = 0/1 = 0]. None of the remaining principal components (PC_2 – PC_{17}) identified correctly the 30 original mice. **Figure 5** illustrates the scores of our 30 original mice with respect to the first and second principal components (PC_1 and PC_2) of the standard PCA with the correlation matrix. One can see from **Figure 5** that there is a separation of the two groups (WT & R6/2) only with respect to PC_1. Table S7 in the Supplementary Material shows the corresponding eigenvectors of the first five principal components; and as can be seen from there, Gln has the greatest weight (0.3577), and it is, therefore, the most significant variable for Factor 1. Observing the absolute value of the weights of the variables, one can see that, in a descending order of significance, the most significant variables for Factor 1 are: 1) Gln, 2) GPC+PC, 3) Cr+PCr, 4) PCr, 5) NAA, 6) Tau, 7) GSH, 8) mIns, 9) PE, 10) Cr, 11) Glu, 12) NAAG, 13) PCr/Cr, 14) Lac, 15) GABA, 16) Glc, and 17) MM. This constitutes a large improvement on the part of the standard PCA (correlation matrix) with respect to the order of the significance of the variables as compared with ROC curve analysis. In other words, using standard PCA with the correlation matrix was a considerable improvement over standard PCA with the covariance matrix. Focusing on the top 10 most important IVs, one can see that they are the same as those identified by ROC curve analysis (**Table 2**). That is, however, the whole extent of the commonality between the two methods. The order of significance of the top ten IVs according to the standard PCA (correlation matrix) is markedly different from that of ROC curve analysis. The most notable differences in that order are the following: 1) Cr+PCr, which has a perfect AUC (1.0000), is placed third and not first; 2) Cr, which has an AUC = 0.9883, almost the same as Gln, is placed tenth and not third; 3) PCr, which has an AUC = 0.8702, is placed fourth (instead of eighth) and ahead of NAA, which has an AUC = 0.9820; 4) Tau, which has the lowest AUC of all ten metabolites (AUC = 0.7289) is placed sixth (instead of tenth) and ahead of NAA.

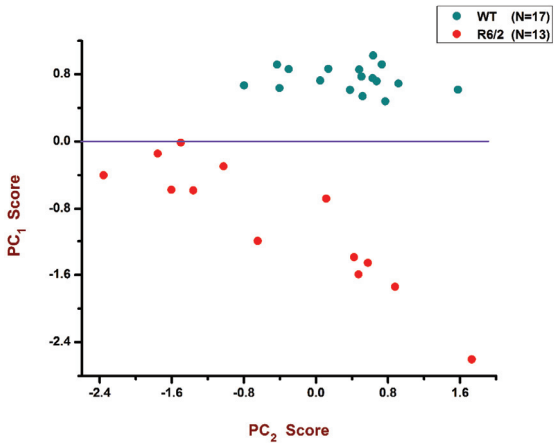

Figure 5. Standard PCA (correlation matrix) – Test 1. Scores of the 30 original mice according to the first principal component (PC_1 Score) plotted against the scores of the same mice according to the second principal component (PC_2 Score). The PCA was run unsupervised (all 17 IVs were used) using the correlation matrix. As can be seen, there is a separation between the two groups [WT (#1-17) & R6/2 (#18-30)] only with respect to the first principal component: all of the WT mice have positive scores, whereas all of the R6/2 mice have negative scores.

Test 2: Identification of the 31 unknown mice (WT vs. R6/2) – Validation Test: We subjected the standard PCA (all 17 IVs) with the correlation matrix to the second and most stringent test, namely the identification of the 31 unknown mice. It correctly identified 28/31 unknown mice [20/20 WT mice (100% correct) and 8/11 R6/2 mice (72.73% correct), with a total accuracy of 28/31 unknown mice (90.32% correct)] [sensitivity = 0.727; (1-specificity) = 0; (+)LR = 0.727/0 $\rightarrow \infty$; (-)LR = 0.273]. Those results in detail, along with the results of all PCA runs with respect to the second test, are shown in **Table 5**. The general results of all PCA runs, including those of this run, appear in **Table 3**.

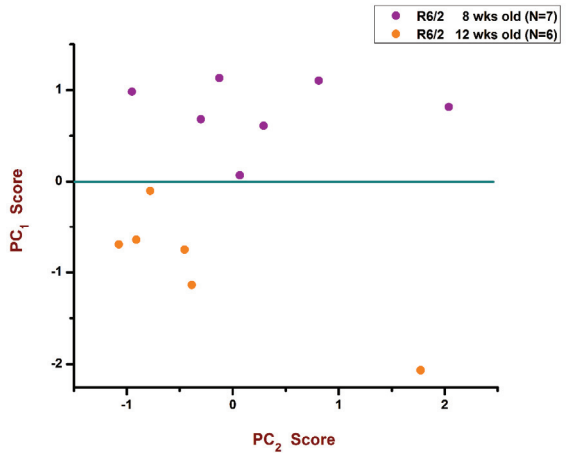

Figure 6. Standard PCA (correlation matrix) – Test 3. Scores of the 13 original R6/2 mice according to the first principal component (PC_1 Score) plotted against the scores of the same mice according to the second principal component (PC_2 Score). The PCA was run unsupervised (all 17 IVs were used) using the correlation matrix. As can be seen, there is a separation between the two groups [8 wk-old R6/2 (#1-7) and 12 wk-old R6/2 (#8-13)] only with respect to the first principal component: all of the 8 wk-old R6/2 mice have positive scores, whereas all of the 12 wk-old R6/2 mice have negative scores.

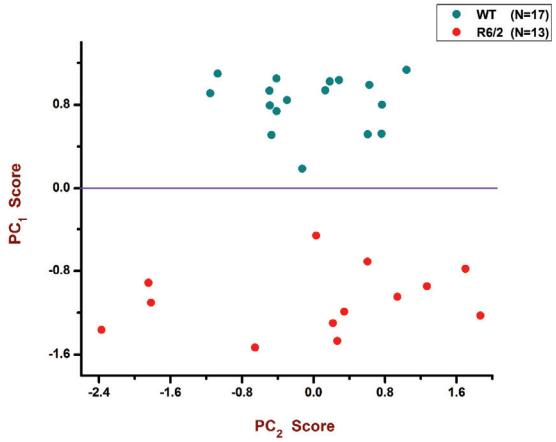

Figure 7. ROC-supervised PCA (correlation matrix) – Test 1. Scores of the 30 original mice according to the first principal component (PC_1 Score) plotted against the scores of the same mice according to the second principal component (PC_2 Score). The PCA was run using the corrrelation matrix and was supervised by the ROC curve analysis [the top four most significant IVs (Cr+PCr, Gln, Cr, and NAA) (AUC > 95%) as determined by the ROC curve analysis were used]. As can be seen, there is a separation between the two groups [WT (#1-17) & R6/2 (#18-30)] only with respect to the first principal component: all of the WT mice have positive scores, whereas all of the R6/2 mice have negative scores.

Test 3: Identification of the 13 original R6/2 mice (8-wk old vs. 12-wk old): Standard PCA (correlation matrix) correctly identified and classified all of our original 13 R6/2 mice into their respective two subgroups [7/7 8 wk-old R6/2 mice (100% correct) & 6/6 12 wk-old R6/2 mice (100% correct) $\rightarrow$ with a total accuracy of 13/13 original R6/2 mice (100% correct)] [sensitivity = 1; (1-specificity) = 0; (+)LR = 1/0 $\rightarrow \infty$; (-)LR = 0/1 = 0]. As can be seen from **Figure 6**, there is a separation of the two R6/2 groups with respect to PC_1: all of the 8 wk-old R6/2 mice have positive scores, whereas all of the 12 wk-old R6/2 mice have negative scores.

The results of this run, along with the general results of all PCA runs, appear in **Table 3**.

ROC-supervised PCA with correlation matrix

Test 1: Identification of the original 30 mice (WT vs. R6/2): Just as we did in the case of the ROC-supervised PCA (covariance matrix), we ran the ROC-supervised PCA (correlation matrix) with the top 10 IVs (AUC > 70%), top 9 IVs (AUC > 80%), top 7 IVs (AUC $\geq$ 90%), and top 4 IVs (AUC > 95%) according to ROC curve analysis (**Table 2**). All four of those runs correctly identified all of the 30 original mice [17/17 WT mice (100% correct) & 13/13 R6/2 mice (100% correct) $\rightarrow$ with a total accuracy of 30/30 original mice (100% correct)]. According to the residuals, the run with the top 10 IVs (AUC > 70%) yielded a sum of all Q_1 residuals equal to 101.42, and a mean Q_1 residual value of 3.38. The respective values of the run with the top 9 IVs (AUC > 80%) were: 85.17 and 2.84. The respective values of the run with the top 7 IVs (AUC $\geq$ 90%) were: 54.72 and 1.82; and those of the run with the top 4 IVs (AUC > 95%) were: 18.02 and 0.60 respectively. Clearly, here, too, the run with the top 4 IVs (AUC > 95%) was the best ROC-supervised PCA setting (correlation matrix) for the 30 original mice, and it was upon the equation of the first principal component of this setting that the ROC-supervised PCA DBM (correlation matrix) was based. As was mentioned, the ROC-supervised PCA DBM

(correlation matrix) correctly identified all of our 30 original mice [17/17 WT mice (100% correct) & 13/13 R6/2 mice (100% correct) → with a total accuracy of 30/30 original mice (100% correct)] [sensitivity = 1; (1-specificity) = 0; (+)LR = 1/0 → ∞; (-)LR = 0/1 = 0]. **Figure 7** illustrates those results. As can be seen, our two groups (WT & R6/2) were successfully separated by the first principal component: all of the WT mice have positive scores, whereas all of the R6/2 mice have negative scores. The general results of all PCA runs, including those of this run, appear in **Table 3**.

Since both the standard PCA (correlation matrix) and the ROC-supervised PCA (correlation matrix) passed the first test, i.e. correctly identified all of the 30 original mice, we compared their respective residuals. In the case of the former, the sum of all Q_1 residuals was 291.11 and the mean Q_1 residual value was 9.70. In the case of the latter [top 4 IVs (AUC > 95%)], the respective values, as already reported above, were: 18.02 and 0.60. Evidently, there is a vast difference in classification accuracy between the standard PCA (correlation matrix) and the ROC-supervised (correlation matrix), albeit both passed the first test.

Test 2: Identification of the 31 unknown mice (WT vs. R6/2) – Validation Test: We subjected the ROC-supervised PCA (correlation matrix) [using only the top 4 IVs (AUC > 98%) according to ROC curve analysis] to the second test. It correctly determined the status of all of the 31 unknown mice [20/20 WT mice (100% correct) and 11/11 R6/2 mice (100% correct), with a total accuracy of 31/31 unknown mice (100% correct)] [sensitivity = 1; (1-specificity) = 0; (+) LR = 1/0 → ∞; (-)LR = 0/1 = 0]. Those results in detail, along with the results of all PCA runs with respect to the second test, are shown in **Table 5**. The general results of all PCA runs, including those of this run, appear in **Table 3**.

Test 3: Identification of the 13 original R6/2 mice (8-wk old vs. 12-wk old): Just as was the case with the R6/2-ROC-supervised PCA (covariance matrix), the best R6/2-ROC-supervised PCA (correlation matrix) setting for the 13 original R6/2 mice both in terms of classification accuracy and residuals was the one that employed only the top two most significant IVs (AUC > 95%), i.e. TTau and GPC+PC; and it is this setting that we used for the third test. This

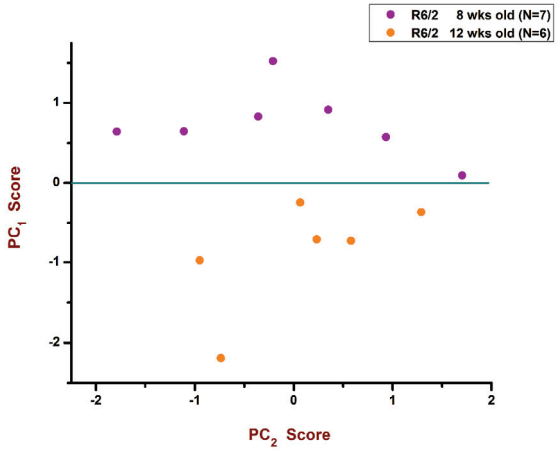

Figure 8. R6/2-ROC-supervised PCA (correlation matrix) – Test 3. Scores of the 13 original R6/2 mice according to the first principal component (PC_1 Score) plotted against the scores of the same mice according to the second principal component (PC_2 Score). The PCA was run using the correlation matrix and was supervised by the R6/2-ROC curve analysis [the top two most significant IVs (TTau and GPC+PC) (AUC > 95%) as determined by the R6/2-ROC curve analysis were used]. As can be seen, there is a separation between the two groups [8 wk-old R6/2 (#1-7) and 12 wk-old R6/2 (#8-13)] only with respect to the first principal component: all of the 8 wk-old R6/2 mice have positive scores, whereas all of the 12 wk-old R6/2 mice have negative scores.

R6/2-ROC-supervised PCA (correlation matrix) correctly identified and classified all of the 13 original R6/2 mice into their respective two subgroups [7/7 8 wk-old R6/2 mice (100% correct) & 6/6 12 wk-old R6/2 mice (100% correct) → with a total accuracy of 13/13 original R6/2 mice (100% correct)] [sensitivity = 1; (1-specificity) = 0; (+)LR = 1/0 → ∞; (-)LR = 0/1 = 0]. **Figure 8** depicts those results. As can be seen, our two R6/2 groups were successfully separated (correctly identified) by the first principal component: all of the 8 wk-old R6/2 mice have positive scores, whereas all of the 12 wk-old R6/2 mice have negative scores. These results are also shown, along with the general results of all PCA runs, in **Table 3**. The numerical results of all PCA runs not presented here can be found in the Supplementary Material.

Since both the standard PCA (correlation matrix) and the R6/2-ROC-supervised PCA (correlation matrix) passed the third test, i.e. correctly identi-

fied all of the 13 original R6/2 mice, we compared their respective residuals. In the case of the former, the sum of all Q_1 residuals was 118.43 and the mean Q_1 residual value was 9.11. In the case of the latter, the respective values were: 1.44 and 0.11. Therefore, in the case of the third test, as well, there was a vast difference in classification and predictive performance between the standard PCA (correlation matrix) and the R/62-ROC-supervised PCA, even though both passed the third test.

ROC-supervised PCA with covariance matrix vs. ROC-supervised PCA with correlation matrix

Finally, given that both ROC-supervised PCAs (covariance and correlation matrix) passed all three tests with 100% accuracy, we wanted to know if there were any differences in the classification and predictive performance of those two methods.

In connection with the first test, the ROC-supervised PCA (covariance matrix) yielded the following: Sum of all Q_1 residuals = 23.11 and Mean value of all Q_1 residuals = 0.77. The ROC-supervised PCA (correlation matrix) yielded respectively: 18.02 and 0.60. This suggests that, all things being equal, the ROC-supervised PCA with the correlation matrix performs better than the ROC-supervised PCA with the covariance matrix in terms of classification and predictive capabilities.

In connection with the third test, the R6/2-ROC-supervised PCA (covariance matrix) yielded the following: Sum of all Q_1 residuals = 11.27 and Mean value of all Q_1 residuals = 0.87. The R6/2 -ROC-supervised PCA (correlation matrix) yielded respectively: 1.44 and 0.11. In the case of the third test, also, the ROC-supervised PCA with the correlation matrix turned out to be more robust than the ROC-supervised PCA with the covariance matrix in terms of classification capabilities.

That the ROC-supervised PCA with the correlation matrix has better classification capability than the ROC-supervised PCA with the covariance matrix is further supported by the following theoretical observations. In the case of the ROC-supervised PCA with the covariance matrix, according to the absolute value of the weights of the variables within the eigenvector of the first principal component (PC_1), the rank of significance of the 4 IVs, including the absolute value of their respective weights, is: 1) Gln [0.6277], 2) Cr+PCr [0.5971], 3) Cr [0.3763], and 4) NAA [0.3284]. According to the ROC curve analysis, the rank of significance of those 4 IVs according to their respective AUC value is: 1) Cr+PCr, 2) Gln, 3) Cr, and 4) NAA (**Table 2**). As we mentioned earlier, and as can also be seen from **Table 2**, Cr+PCr is the only perfect biomarker (AUC = 1.0000), and it is upon it that PC_1 should be based. Similarly, in the case of the ROC-supervised PCA with the correlation matrix, the rank of significance of the 4 IVs, including the absolute value of their respective weights, is: 1) Cr+PCr [0.5303], 2) Gln [0.4993], 3) NAA [0.4852], 4) Cr [0.4838]. This shows that the PC_1 of the ROC-supervised PCA with the correlation matrix was based predominantly on the Cr+PCr variable, which has a perfect discriminating power (AUC = 1.0000). That further indicates that the ROC-supervised PCA with the correlation matrix has a better classification and predictive capability than the ROC-supervised PCA with the covariance matrix.

Discussion

The results of our study (**Table 3**) demonstrate that our ROC-supervised PCA may be employed for the diagnosis of diseases. More specifically, both ROC-supervised PCA with the covariance matrix and ROC-supervised PCA with the correlation matrix passed all three stringent tests with 100% accuracy, exhibiting, thus, high diagnostic accuracy, and providing evidence that they may be used for diagnostic purposes.

The fact that both of those methods yielded results that were 100% accurate notwithstanding, as was demonstrated in the previous section, the ROC-supervised PCA with the correlation matrix exhibited a better classification and predictive capability than the ROC-supervised PCA with the covariance matrix.

Standard PCA, on the other hand, be it with the covariance or the correlation matrix, did not pass all of our three tests (**Table 3**), and that provides evidence against its employment in diagnostic applications. More specifically, standard PCA (covariance matrix) failed all three of our tests, thus proving itself unsuitable for diagnostic applications; whereas standard PCA (correlation matrix) passed the first and the

third test but failed the second test (the validation test, i.e. the most difficult of the three tests), thus demonstrating that it lacks the high degree of accuracy required for the diagnosis of diseases. The primary objective of the standard PCA algorithm is to reduce the dimensionality of the data by seeking to maximize the amount of the original variance (information) in the direction of the variable(s) with the largest variance. Unfortunately, the largest variance, the largest amount of information, is not always synonymous with the most significant information. Commenting on this issue, Mather [7] pointed out that "A major problem in PCA is the distinction between important and unimportant dimensions of variability." The results of our study clearly support this contention. Therefore, employing standard PCA for diagnostic or other purposes requiring a high degree of accuracy may not constitute a wise choice. Today, PCA has found its way in the main stream of biomedical research. Owing to significant advances of technology, such as the ability to gather information about large numbers of metabolites, genes, or proteins, vast amounts of data can be generated. Confronted by such a plethora of data, researchers have little choice but to resort to data analysis/mining methods, such as PCA. On account of many reasons, including ease of use, PCA, in one form or another, has become popular. Many researchers routinely entrust their data to PCA and predicate their study conclusions on the results yielded by it [15-19].

As we have shown, our ROC-supervised PCA, especially with the correlation matrix, possesses the high degree of classification and predictive accuracy that is prerequisite in the diagnosis of diseases. We should also point out here that insofar as accuracy and performance are concerned, according to the results of our previous studies, our ROC-supervised PCA provides a competitive alternative to other more complex multivariate methods [14], as well as other data analysis methods [20]. There is a limitation, however, that underlies its applicability. Owing to the fact that the outcome of the dependent variable in ROC curve analysis is dichotomous (only two outcomes are possible, i.e. WT or R6/2 in our case), and since our PCA is supervised by ROC curve analysis, it follows that our ROC-supervised PCA can be applied only to those diseases wherein there are only two groups (or two classifications). This, however, in actuality, may not be as restrictive as it sounds for the following two reasons. As it turns out, in most of the disease states, researchers are, at least initially, interested in differences between the state of a given disease and the normal state. Secondly, if in a given disease a researcher is indeed interested in three groups, let us say, normal, pre-symptomatic, and pathological, then three different ROC curve analyses (one between normal and pre-symptomatic, one between pre-symptomatic and pathological, and one between normal and pathological) may be performed and used with our ROC-supervised PCA. However, if the number of groups (or classifications) is greater than three, then this approach may be unrealistic.

Furthermore, we should point out that owing to the fact that the spectrum of diseases and disorders is very wide and variegated, the degree of accuracy will vary in accordance with the specific conditions of the particular disease and with the desired type of diagnostic model. For instance, if one is interested in colorectal cancer (CRC), and if, furthermore, one is interested in the differential diagnosis between normal subjects and patients with stage II CRC because the majority of the CRC patients when first diagnosed present with stage II, then the degree of accuracy of ROC-supervised PCA will be higher since the contrast between the normal and the specific diseased state is relatively large. In the case of those diseases where the patient population is not as homogeneous in terms of severity, extent of impairment, progression, symptomatology, etc., finding a suitable reference point for a diagnostic model will undoubtedly be more challenging, and the performance of ROC-supervised PCA in that case will be dependent on the careful selection of the variables, as well as on larger sample sizes.

We should also point out here that although the idea of seeking to improve the standard PCA is not new, our PCA method is, as far as we know, novel and fundamentally different from those proposed by others. For example, Bair et al. [22] proposed a supervised PCA method based on standard regression and applied to survival analysis. Nguyen and Rocke [23] and Hi and Gui [24] proposed various PLS (partial least squares) methods also in connection with survival predictions. Our ROC-supervised PCA was developed specifically for diagnostic purposes, and it is predicated on the screening and selection of data variables according to their discrimi-

nating accuracy between the target and the reference group. Moreover, unlike in the case of other proposed supervised PCA methods, our ROC-supervised PCA considers only the first principal component, which as we have shown above captures most of the variance of the original variables and has the highest potential in terms of classification accuracy than any other principal component.

In conclusion, in the present study, we assessed the performance and brought to light the weaknesses of standard PCA in connection with classification accuracy and diagnostics; we introduced our ROC-supervised PCA that was developed to address specifically the weaknesses of standard PCA in that area; we assessed the classification and predictive accuracy of our ROC-supervised PCA and compared it to that of the standard PCA; and we provided evidence that supports the use of our ROC-supervised PCA for diagnostic purposes.

Acknowledgements

We would like to thank C. Dirk Keene and Ivan Tkac for helping us with the acquisition of spectra and Janet M. Dubinsky for providing us with the spectral data of 20 unknown mice. This study was funded by the National Institutes of Health (NIH) - Grant numbers: T32 DA007097 and R03 NS060059.

Please address correspondence to: Dr. Jason B. Nikas, Department of Neurosurgery, Medical School of University of Minnesota, 4-218 MTRF, 2001 Sixth St., SE, Minneapolis, MN 55455, USA. Tel: 612-625-2868, Fax: 612-626-9201, E-mail: nikas001@umn.edu

References

[1] Pearson K. On lines and planes of closest fit to systems of points in space. Philosophy 1901; 2: 559-572.

[2] Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 1933; 24: 417-441.

[3] Dunteman GH. Principal Components Analysis. Newbury Park, CA: Sage University Paper series on Quantitative Applications in the Social Sciences, No 07-069; 1989.

[4] Jackson JE. A User's Guide to Principal Components. New York, NY: John Wiley & Sons; 1991.

[5] Jolliffe IT. Principal Component Analysis. New York, NY: Springer-Verlag; 2002.

[6] McArdle JJ. Principles versus Principals of Structural Factor Analysis. Multivariate Behavioral Research 1990; 25: 81-87.

[7] Mather PM. Computational Methods of Multivariate Analysis in Physical Geography. London: John Wiley & Sons; 1976.

[8] Costello AB, Osborne JW. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. Practical Assessment Research & Evaluation 2005; 10: 7.

[9] Velicer WF, Jackson DN. Component Analysis versus Common Factor Analysis: Some Issues in Selecting an Appropriate Procedure. Multivariate Behavioral Research 1990; 25: 1-28.

[10] Mangiarini L, Sathasivam K, Seller M, Cozens B, Harper A, Hetherington LM, Trottier Y, Lehrach H, Davies SW, Bates GP. Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. Cell 1996; 87: 493–506.

[11] Tkac I, Henry PG, Andersen P, Keene CD, Low WC, Gruetter R. Highly resolved in vivo 1H NMR spectroscopy of the mouse brain at 9.4 T. Magn. Reson. Med. 2004; 52: 478–484.

[12] Browne SE, Beal MF. The Energetics of Huntington's Disease. Neurochemical Research 2004; 29: 531–546.

[13] Tkac I, Dubinsky JM, Keene CD, Gruetter R, Low WC. Neurochemical changes in Huntington R6/2 mouse striatum detected by in vivo 1H NMR spectroscopy. Journal of Neurochemistry 2007; 100: 1397-406.

[14] Nikas JB, Keene CD, Low WC. Comparison of Analytical Mathematical Approaches for Identifying Key Nuclear Magnetic Resonance Spectroscopy Biomarkers in the Diagnosis and Assessment of Clinical Change of Diseases. Journal of Comparative Neurology 2010; 518: 4091–4112.

[15] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 2006; 38: 904-909.

[16] Gelson W, Hoare M, Unitt E, Palmer C, Gibbs P, Coleman N, Davies S, Alexander GJM. Heterogeneous Inflammatory Changes in Liver Graft Recipients With Normal Biochemistry. Transplantation 2010; 89: 739–748.

[17] Hillegass JM, Shukla A, Macpherson MB, Bond JP, Steele C, Mossman BT. Utilization of gene profiling and proteomics to determine mineral pathogenicity in a human mesothelial cell line (LP9/TERT-1). J Toxicol Environ Health A. 2010; 73: 423–436.

[18] Rohrbeck A, Borlak J. Cancer Genomics Identifies Regulatory Gene Networks Associated with the Transition from Dysplasia to Advanced Lung Adenocarcinomas Induced by c-Raf-1. PLoS ONE 2009; 4(10) e7315. doi:10.1371/journal.pone.0007315

[19] Massad LS, Evans CT, Wilson TE, Goderre JL,

Hessol NA, Henry D, Colie C, Strickler HD, Levine AM, Watts DH, Weber KM. Knowledge of cervical cancer prevention and human papillomavirus among women with HIV. Gynecologic Oncology 2010; 117: 70-76.

[20] Nikas JB, Low WC. Clustering Analyses for the Diagnosis of Huntington and Other Diseases (Abstracts for the 17th Annual Meeting of the American Society for Neural Therapy and Repair). Cell Transplantation 2010; 19: 355.

[21] Hintze JL. NCSS 2007 Manual. Kaysville, Utah : NCSS; 2007.

[22] Bair E, Hastie T, Paul D, Tibshirani R. Prediction by Supervised Principal Components. Journal of the American Statistical Association 2006; 101: 119-137.

[23] Nguyen D, Rocke D. Partial Least Squares Proportional Hazard Regression for Application to DNA Microarrays. Bioinformatics 2002; 18: 1625–1632.

[24] Li H, Gui J. Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data. Bioinformatics 2004; 20: (Suppl 1) i208–i215.

Supplementary Material for AJTR1101001
Am J Transl Res 2011;3(2):180-196.

ROC-supervised Principal Component Analysis in connection with the diagnosis of diseases

Jason B. Nikas[1,2], Walter C. Low[1,3,4,5,6]

[1]Department of Neurosurgery, [2]Pharmaco-Neuro-Immunology Program, [3]Graduate Program in Neuroscience, [4]Department of Integrative Biology and Physiology, [5]Institute for Translational Neuroscience, [6]Center for Neuroengineering, Medical School, University of Minnesota, Minneapolis, MN, USA

*Corresponding Author: Jason B. Nikas
Department of Neurosurgery, Medical School
University of Minnesota
4-218 MTRF
2001 Sixth St., SE
Minneapolis, MN 55455, USA
T: 612-625-2868 / F: 612-626-9201
E-mail: nikas001@umn.edu

1. BASIC PCA THEORY

To illustrate the way PCA functions, let us consider the following example. Given two variables, X_1 and X_2, and assuming that that each distribution of those two variables has a mean value equal to zero, i.e. the mean of each variable has been subtracted from every data point of that variable, one would see, if one plotted X_1 variable vs. X_2 variable, that all data points are contained in a space very much resembling that enclosed by the blue-outlined ellipse in Figure S1 (assuming that our data are normally distributed and correlated to some extent) [3,7]. First, PCA will compare the variances of X_1 and X_2. From Figure S1, the variance of X_1 is A'B', while the variance of X_2 is A"B". We can also see that A'B' > A"B". Since, from our two variables, X_1 has the greater variance, PCA will zero in on it and will seek to maximize its variance; or equivalently, PCA will seek to maximize the variance of all of our data points. Looking at Figure S1, one can conclude that the only way that that can be achieved is by drawing a straight line along the major axis of the ellipse. Let us call that line PC_1. Then, with respect to line PC_1, the variance of our data points is AB. One can easily verify that AB > A'B'. The line PC_1 is unique: it is the only line that causes the sum of the squared distances of all of our points from it to be the smallest. In other words, PC_1 is the line of best (closest) fit to our points. It is called Principal Component 1 (or sometimes Factor 1), and it is the most important of all the principal components in that it has the largest variance, i.e. the largest amount of information that was contained in both of our original variables (X_1 and X_2). PC_1 can be expressed in terms of X_1 and X_2.

$$PC_{1N} = w_{11}X_{1N} + w_{12}X_{2N} \quad (1.1)$$

or in the case of P variables:

$$PC_{1N} = w_{11}X_{1N} + w_{12}X_{2N} + \ldots + w_{1P}X_{PN} \quad (1.2)$$

X_{1N}, X_{2N}, ..., X_{PN} are the P variables (in our case, P=17 metabolite concentrations) of subject N; w_{11}, w_{12}, ... w_{1P} are the weights of the P variables with respect to PC_1; and PC_{1N} is the score of subject N with respect to the first principal component (PC_1).

The weights w_{11} and w_{12} are chosen in such a way that the variance of PC_1 is maximized subject to the following condition:

$$(w_{11})^2 + (w_{12})^2 = 1 \quad (1.3)$$

or in the case of P variables:

$$(w_{11})^2 + (w_{12})^2 + \ldots + (w_{1P})^2 = 1 \quad (1.4)$$

Owing to the fact that, as we saw earlier, the variance of X_1 is larger than that of X_2, the weight of X_1 will be larger than that of X_2: $w_{11} > w_{12}$. If the variance of X_1 is much greater than that of X_2, then $w_{11} >> w_{12}$. The weights, therefore, in Equations (1.1) and (1.2) are proportional to and indicative of the magnitude of the variance, or, equivalently, the magnitude of significance that each of the original variables has for the given principal component. We should point out here that if, instead, we had P variables, with P>>2, and if a small number of the those variables had variances much larger than the rest, then from Equations (1.2) and (1.4), one can see that a small number of the weights (w_{11}, w_{12}, w_{13}, ... w_{1P}) would be significant, and the rest of them would be close to zero. That also means that a small number of the original variables (X_1, X_2, X_3, ... X_P) would be significant as far as PC_1 is concerned; whereas the rest of them would not be so. In that case, one may elect to keep that small number of variables and discard the rest of them, thus reducing considerably the number of the original variables [7].

As stated above, the first principal component (PC_1) is the most important of all principal components because it contains most of the information (variance) of the original variables [3-5,7]. The second principal component (PC_2) is the second most significant because it contains the second largest amount of the original variance, and it follows that the first few principal components are significant, whereas the remaining ones are not.

Using matrix theory, we can express the weights of PC_1 and PC_2 into a vector form.

$$\mathbf{W}_1 = [w_{11}, w_{12}, \ldots w_{1P}] \quad (1.5)$$

$$\mathbf{W}_2 = [w_{21}, w_{22}, \ldots w_{2P}] \quad (1.6)$$

Equation (1.5) gives us the eigenvector of PC_1, while Equation (1.6) gives us the eigenvector of PC_2. As we mentioned earlier, the variance of PC_1 is AB, and the variance of PC_2 is CD (Figure S1). Let us say that AB = λ_1 and CD = λ_2. The set

$$\mathbf{\Lambda} = [\lambda_1, \lambda_2] \quad (1.7)$$

gives us the eigenvalues of the eigenvectors $\mathbf{W}_1$ and $\mathbf{W}_2$ respectively [Equations (1.5) and (1.6)]. In other words, the variance of a principal component is the eigenvalue of the eigenvector of that component. In the case of P variables, Equation (1.7) becomes:

$$\mathbf{\Lambda} = [\lambda_1, \lambda_2, \ldots \lambda_P] \quad (1.8)$$

Equation (1.8) gives us the eigenvalues of all P eigenvectors $\mathbf{W}_1$, $\mathbf{W}_2$,...,$\mathbf{W}_P$.

In terms of matrix notation, therefore, PCA starts from the following:

$$\mathbf{Y} = \mathbf{WX} \quad (1.9)$$

Since the vectors (lines) of all principal components go through the origin (0,0), Equation (1.9) is the general form of the vectors of all principal components in terms of all the variables (X). The slope of all those vectors is the eigenvector $\mathbf{W}$, which is what PCA seeks to calculate. Next, PCA will form the variance-covariance matrix, $\mathbf{S}$, of all the variables. As we discussed earlier, all variables are transformed so their mean is equal to zero. If $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues of $\mathbf{S}$, and if $\mathbf{W}^*$ is the matrix of the eigenvectors of $\mathbf{S}$, then we have the final solution:

$$\mathbf{W} = \mathbf{W}^* \mathbf{\Lambda}^{-1/2} \quad (1.10)$$

From Equation (1.10) PCA calculates all eigenvectors, i.e. the weights (coefficients) of the lines of all principal components. We should mention here that Equations (1.9) and (1.10) yield standardized principal components, i.e. the mean of the weights of each principal component is equal to zero and the standard deviation of the weights of each principal component is equal to one [21].

2. PCA WITH COVARIANCE MATRIX VS. PCA WITH CORRELATION MATRIX

The very first step of PCA using the covariance matrix is to transform all IVs so that each one of them has a mean value equal to zero. If, in addition to that, we were to divide every observation (data point) within each IV by that IV's standard deviation, then all of the IVs would have a mean value equal to zero and also a standard

deviation equal to one. That is the definition of a standardized variable. Since variance is equal to standard deviation squared, it follows that all standardized IVs have a variance that is also equal to one. If we standardized all of our IVs, and then formed the covariance matrix, that matrix would be the correlation matrix, and that is how PCA forms the correlation matrix should one choose to run the PCA thus. What are the differences between the two methods? Technically, there is only one difference. Since every IV has a variance of one, and since, let us say, we have P IVs, the total variance of all of our variables is equal to P. This means that the sum of the eigenvalues of all P principal components [from Equation (1.8)] is equal to P:

$$\lambda_1 + \lambda_2 + \ldots + \lambda_P = P \quad (1.11)$$

In the case of the covariance matrix (non-standardized IVs), the sum of the eigenvalues of all P principal components is equal to the sum of the variances of all P principal components:

$$\lambda_1 + \lambda_2 + \ldots + \lambda_P = \lambda_T \quad (1.12)$$

(where in all likelihood, $\lambda_T \neq P$)

We should point out here that by standardizing all IVs, i.e. by making their variance equal to one, in effect, what we are doing is to reduce significantly any original large variance disparities among our IVs.

3. POTENTIAL PROBLEMS FOR PCA

To provide a better understanding of the mechanism of function of PCA, we can avail ourselves of the biomarker data and results of this study. Let us return to our previous example in the first section (Basic PCA Theory) and make it more specific by applying it to our data and results. Let us say that variable X_1 is the striatal concentrations of the metabolite Tau of 17 WT and 13 R6/2 mice, and that variable X_2 is the striatal concentrations of the metabolite Cr+PCr of the same 17 WT and 13 R6/2 mice. Figure S2 provides a diagrammatical representation of the two aforementioned variables. We note that the variance of Tau is 5.163, while the variance of Cr+PCr is 1.841. We also note from Figure S2 that in the case of Tau, there is a significant amount of overlapping with respect to our two groups (WT & R6/2); whereas in the case of Cr+PCr, there is no overlapping. That is to say, the ratio of the between-group variance over the within-group variance is in fact much larger for Cr+PCr than it is for Tau. Obviously, if one is interested in finding accurate biomarkers that can be used to accurately diagnose an unknown mouse or patient in terms of HD, then Cr+PCr is the significant biomarker, whereas Tau is not. For diagnostic or clinical change assessment purposes, Cr+PCr is the ideal biomarker, whereas Tau is not a significant biomarker.

If we perform PCA on our two variables (Tau and Cr+PCr), owing to the fact that Tau has a variance that is almost three times larger than that of Cr+PCr, PCA will lock onto Tau and will seek to maximize its variance by fitting the best line (PC_1). What we described earlier in our example with the two variables (X_1 and X_2) is exactly what will transpire now that we replaced X_1 with Tau and X_2 with Cr+PCr. Since Tau has a significantly larger variance than Cr+PCr, its weight (coefficient) will also be significantly larger than that of Cr+PCr. That means that the line equation of the first principal component will be heavily influenced by Tau as opposed to Cr+PCr. That ultimately means that a diagnosis, given by the equation of the first principal component, will not be accurate. In fact, from our ROC curve analysis (Table 2), one can see that Tau has an AUC = 0.7289, which means that its diagnostic power as a biomarker is 72.89%, whereas Cr+PCr has an AUC = 1.0000, which is perfect, and which means that its diagnostic power as a biomarker is 100%.

SUPPLEMENTARY TABLES

Table S1. Classification of the 30 original mice (17 WT & 13 R6/2) by the first six principal components (factors) of standard PCA using the covariance matrix. All 17 IVs were used in this run. Rows #1-17 are the WT mice, rows #18-30 are the R6/2 mice. None of the 17 factors (the first 6 of which are shown here) identified correctly all of the 30 mice. The first principal component (Factor 1) misidentified 5 mice (#19-22 & #24), which are R6/2, and which should have negative factor scores. The rest of the factors (2-17) did not show any meaningful results with respect to the identification of the 30 original mice.

Factor Score

	Factors					
Row	**Factor1**	**Factor2**	**Factor3**	**Factor4**	**Factor5**	**Factor6**
1	0.6749	0.3897	1.6938	1.8833	1.2820	-0.3479
2	0.6117	-0.4154	-0.3352	-1.3478	1.2889	1.8374
3	0.7297	-0.5802	1.6604	0.1681	-0.2911	0.7011
4	0.4323	0.2930	1.3839	1.3065	0.8591	0.1816
5	0.4535	1.3081	-0.1864	-0.0651	0.8633	-0.3314
6	0.7363	0.6956	-0.4075	-1.0891	1.2536	0.7100
7	0.5430	1.1575	-0.8889	0.1418	1.0074	-0.0137
8	0.5104	0.7138	0.4588	-0.8361	0.1870	0.2482
9	0.4466	0.9394	0.1507	-0.5127	0.1964	-0.5563
10	0.5184	1.1827	-0.5560	-1.1121	-1.5475	-0.3809
11	0.6183	0.1019	1.1848	-0.8550	-1.5236	-1.5284
12	0.7028	0.4429	-0.2723	0.7200	-0.8357	0.0148
13	0.3978	0.6855	-0.5649	-0.4905	0.1125	-1.3124
14	0.7652	0.1439	-0.3013	-0.1833	-0.3952	0.1364
15	0.7421	0.7372	-0.3556	0.8057	-0.2820	0.6563
16	0.8077	0.4460	0.5932	1.0454	-1.1802	0.7963
17	0.7993	0.6654	-0.0892	0.2580	-0.8904	-0.4079
18	-0.6552	-0.1012	0.3342	-1.0338	0.5643	-1.1647
19	0.0049	-1.0111	-1.5552	2.1456	-1.8030	-0.5345
20	0.2323	-1.5772	-0.4209	-0.7577	0.3122	-0.1218
21	0.0277	-1.7031	-1.1178	-0.4910	0.2340	-1.9802
22	0.2494	-2.4883	1.1260	-1.1081	-0.9489	0.3401
23	-0.2287	-1.7072	1.0513	1.0983	0.6082	0.3994
24	0.4475	-1.6036	-1.0775	-0.8903	1.0185	1.0357
25	-1.0714	-0.4910	-1.3868	0.8777	-1.3363	0.4279
26	-1.5223	0.4812	-0.9421	0.6486	-0.2768	2.9186
27	-1.7734	-0.3715	0.2626	1.5816	1.9764	-1.4694
28	-1.2674	0.0431	0.2314	-0.6079	0.0270	0.2606
29	-1.7479	0.7181	-1.7024	-0.1310	0.6832	-0.7832
30	-3.1855	0.9048	2.0290	-1.1688	-1.1634	0.2681

Table S2. The eigenvalues of the eigenvectors of all 17 principal components (factors) of standard PCA using the covariance matrix. As can be seen, the first principal component (PC_1) accounts for 57.51% of the original variance of the data; the second one (PC_2) accounts for 22.29%; the third one (PC_3) for 8.35%, etc. It's worth noticing that the first four principal components collectively account for 91.53% of the original variance. Since an eigenvalue represents the variance of a principal component, it can be seen that PC_1 has the largest variance (9.621284) of all principal components. All 17 IVs were used in this run.

Principal Components Report

Eigenvalues

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	9.621284	57.51	57.51	|||||||||||
2	3.729402	22.29	79.81	|||||
3	1.397321	8.35	88.16	||
4	0.564596	3.37	91.53	|
5	0.353285	2.11	93.65	|
6	0.260879	1.56	95.21	|
7	0.205806	1.23	96.44	|
8	0.170490	1.02	97.45	|

Table S3. Classification of the 13 original R6/2 mice [(7) 8-wk old & (6) 12-wk old] by the first six principal components (factors) of standard PCA using the covariance matrix. All 17 IVs were used in this run. Rows #1-7 are the 8-wk old R6/2 mice, whereas rows #8-13 are the 12-wk old R6/2 mice. None of the 17 factors (the first 6 of which are shown here) identified correctly all of the 13 R6/2 mice. The first principal component (Factor 1) misidentified 1 mouse (#1), which is an 8-wk old R6/2, and which should have a positive factor score. The rest of the factors (2-17) did not show any meaningful results with respect to the identification of the 13 original R6/2 mice.

Principal Components Report

Factor Score Factors

Row	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
1	-0.0554	0.1709	1.3098	-1.1378	-0.8716	-1.4662
2	0.6812	-1.1919	-1.4989	0.6653	-0.6835	-2.0884
3	0.9509	0.1640	0.4772	-0.4640	-0.1517	-0.0833
4	0.8682	-0.2404	0.1379	-0.3204	2.6479	-0.1309
5	1.1487	1.6523	0.1650	0.9886	0.4098	-0.1077
6	0.6209	1.1182	-1.1678	0.0192	-0.9324	0.8017
7	1.1241	-0.3746	1.1585	-0.3872	-0.7605	1.4609
8	-0.2025	-0.9283	-0.4858	1.1834	0.7824	0.4276
9	-0.7781	-0.9747	0.2074	1.5899	-0.7833	1.1586
10	-0.7600	0.3389	-1.9502	-1.8489	-0.0420	0.7852
11	-0.5156	0.1305	0.7237	0.0809	-0.6224	-0.4754
12	-0.9641	-1.4142	0.5739	-1.0199	0.4812	0.2255
13	-2.1182	1.5495	0.3492	0.6510	0.5261	-0.5076

Table S4. Classification of the 30 original mice (17 WT & 13 R6/2) by the four principal components (factors) of ROC-supervised PCA using the covariance matrix. The top 4 IVs (Cr+PCr, Gln, Cr, and NAA) (AUC > 98%) according to the ROC curve analysis were used in this run. Rows #1-17 are the WT mice, rows #18-30 are the R6/2 mice. The first principal component (Factor 1) correctly identified all of the 30 mice: all WT mice have positive factor scores, whereas all R6/2 mice have negative factor scores. The rest of the factors (2-4) did not show any meaningful results with respect to the identification of the 30 original mice.

Principal Components Report

Factor Score

	Factors			
Row	**Factor1**	**Factor2**	**Factor3**	**Factor4**
1	0.8961	0.5871	-1.2624	-1.2058
2	0.2600	0.0572	0.6988	-1.6878
3	0.5261	0.4128	0.5083	-0.7083
4	0.6339	1.1622	-0.9504	-1.1199
5	1.0018	1.1465	0.0700	0.5141
6	1.0191	0.3325	0.0613	1.2008
7	0.9247	0.2197	-0.2573	-1.4879
8	0.8989	0.0953	0.6866	-0.2377
9	0.9610	-0.4215	-0.7917	1.1142
10	1.0079	0.3928	1.7818	0.9745
11	0.8047	-0.4831	-0.6266	-0.2567
12	0.7115	0.6139	0.0971	-1.2049
13	0.5708	-0.6564	-0.2423	-0.3336
14	0.6593	-0.8952	0.6794	-0.4047
15	1.0479	-0.5904	-1.6115	1.7901
16	1.0815	-0.4598	0.1227	1.1284
17	0.9534	0.2814	0.5530	0.5036
18	-0.5365	0.0981	-0.6422	1.0126
19	-0.9772	-1.0774	1.5113	-0.4023
20	-0.7011	-1.4697	-0.7611	-0.8032
21	-1.0965	-1.8259	-0.4649	-0.7054
22	-0.8828	-1.3816	0.6134	0.0781
23	-0.9882	-0.7562	-1.4056	-0.4659
24	-0.6466	-0.8526	0.0554	0.9186
25	-1.3364	-0.8829	0.9266	0.6655
26	-1.0224	1.0656	2.2195	-0.9440
27	-1.8511	1.7244	-2.1163	-0.9978
28	-0.9649	1.4558	1.1328	0.1336
29	-1.3209	-0.3127	-0.3350	1.1431
30	-1.6343	2.4202	-0.2507	1.7887

Table S5. Classification of the 13 original R6/2 mice [(7) 8-wk old & (6) 12-wk old] by the two principal components (factors) of ROC-supervised PCA using the covariance matrix. The top two most significant IVs (TTau and GPC+PC) (AUC > 95%) according to the R6/2 ROC curve analysis were used in this run. Rows #1-7 are the 8-wk old R6/2 mice, whereas rows #8-13 are the 12-wk old R6/2 mice. The first principal component (Factor 1) correctly identified all of the 13 original R6/2 mice: all of the 8-wk old have negative factor scores, whereas all of the 12-wk old have positive factor scores. The second factor did not show any meaningful results with respect to the identification of the 13 original R6/2 mice.

Principal Components Report

Factor Score

	Factors	
Row	**Factor1**	**Factor2**
1	-0.5062	-1.6298
2	-0.3559	1.2337
3	-0.7174	0.5510
4	-0.1882	1.8905
5	-0.9713	-0.1163
6	-0.7839	-0.7653
7	-1.4264	0.5727
8	0.2221	-0.1237
9	1.1781	0.6839
10	0.6366	-0.3996
11	0.0385	-1.3388
12	0.5709	-0.7407
13	2.3031	0.1823

Table S6. Classification of the 30 original mice (17 WT & 13 R6/2) by the first six principal components (factors) of standard PCA using the correlation matrix. All 17 IVs were used in this setting. Rows #1-17 are the WT mice, rows #18-30 are the R6/2 mice. Of the 17 factors, Factor 1 was the only one to identify correctly all of the 30 mice. As can be seen from the Factor 1 column, all of the WT mice have positive factor scores, whereas all of the R6/2 mice have negative scores. The rest of the factors (2-17) did not show any meaningful results with respect to the identification of the 30 original mice.

Principal Components Report

Factor Score Factors

Row	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
1	0.9175	0.7311	1.1114	-0.4337	1.1843	-2.0022
2	0.6683	-0.7990	0.2473	1.0141	0.8502	1.6412
3	0.6362	-0.4039	1.1697	-0.4664	0.0091	-0.0353
4	0.5397	0.5180	0.9484	-0.5819	0.9880	-0.4563
5	0.6170	1.5764	0.9174	-0.0193	-0.8022	0.4068
6	0.4794	0.7704	0.7345	0.9713	-1.2939	1.1049
7	0.7551	0.6285	-0.9233	0.7616	0.0704	-1.0311
8	0.8583	0.4792	0.0223	1.7817	1.0671	-0.5594
9	0.7185	0.6734	-0.0247	-0.0953	0.2221	-0.2271
10	0.6920	0.9153	-0.2438	-0.6691	-2.2789	1.6880
11	0.8628	-0.3026	-0.3690	-0.3155	0.0591	-0.4926
12	0.8660	0.1388	-0.7693	-1.1109	-0.1997	-0.2079
13	0.7261	0.0475	-0.7473	-0.7503	0.3280	0.6716
14	0.9151	-0.4323	-0.7174	-0.0895	0.7341	0.3116
15	0.7745	0.5061	-0.3913	0.0408	0.3728	-0.4121
16	0.6137	0.3787	-0.2814	1.0151	-0.7147	-1.4435
17	1.0244	0.6349	-0.0790	-1.1598	-0.0088	0.5512
18	-0.6799	0.1137	0.8832	0.4006	-0.6110	1.3479
19	-0.2952	-1.0263	-1.1237	-1.2611	-0.9437	-0.6006
20	-0.1446	-1.7571	0.1100	1.4915	-0.4414	0.4409
21	-0.5746	-1.6054	0.8335	-1.2129	-2.6748	-1.2629
22	-0.4007	-2.3570	1.5440	0.3036	-0.1808	-0.1802
23	-0.5823	-1.3614	0.9365	0.9277	1.1554	-1.1838
24	-0.0121	-1.4990	-0.3014	-0.1280	1.3519	2.1087
25	-1.1949	-0.6463	-1.8695	-1.7268	0.7628	-0.5782
26	-1.5940	0.4720	-1.0863	-0.5308	0.1583	-0.6855
27	-1.7398	0.8784	-1.3850	2.5782	-1.2397	-0.8147
28	-1.3872	0.4231	-0.3946	0.5044	0.1644	0.5821
29	-1.4558	0.5762	-1.3085	-0.1496	1.1107	1.1642
30	-2.6035	1.7287	2.5572	-1.0895	0.8009	0.1546

Table S7. The eigenvectors of the first 5 principal components (factors) of standard PCA using the correlation matrix. The magnitude of the absolute value of the weights of the variables within each eigenvector (column) is directly proportional to the significance of the variables for each factor. As can be seen from the absolute value of the weights of Factor 1 (PC_1), Gln has the greatest weight (0.357688), and it is, therefore, the most significant variable for Factor 1. In a descending order of significance, the most significant variables for Factor 1 are Gln, GPC+PC, Cr+PCr, PCr, NAA, Tau, GSH, mIns, PE, Cr, Glu, NAAG, PCr/Cr, Lac, GABA, Glc, and MM. All 17 IVs were used in this setting.

Principal Components Report

Eigenvectors

	Factors				
Variables	**Factor1**	**Factor2**	**Factor3**	**Factor4**	**Factor5**
Cr	-0.235781	-0.386264	-0.120655	0.003418	-0.020606
Gln	-0.357688	-0.059689	-0.005031	0.048799	0.047574
NAA	0.298498	0.196187	0.139391	0.177390	0.065750
Cr+PCr	-0.342756	-0.190922	-0.036933	0.000722	0.004734
PCr	-0.334808	0.172439	0.097889	-0.003322	0.038440
Glc	0.038925	-0.359440	0.275529	0.148796	-0.163210
Glu	-0.187748	0.338639	0.122520	0.087980	0.205878
GSH	-0.262355	-0.172541	-0.245462	0.268163	0.061304
mIns	-0.251501	-0.217493	0.144344	0.172588	0.376904
Lac	0.110837	-0.327280	0.398691	0.111705	0.073918
PE	0.248826	0.062493	-0.049616	0.192933	0.236967
Tau	-0.298410	0.271671	0.075940	-0.045003	0.172189
GPC+PC	-0.351764	0.104905	-0.019775	-0.112063	-0.034511
MM	0.032920	-0.123482	0.581335	-0.343199	0.434970
GABA	-0.056101	0.124293	0.263922	0.787601	-0.143148
NAAG	-0.157921	0.009027	0.397193	-0.177077	-0.687931
PCr/Cr	-0.113063	0.441599	0.223796	-0.028426	-0.035409

Table S8. Classification of the 13 original R6/2 mice [(7) 8-wk old & (6) 12-wk old] by the first six principal components (factors) of standard PCA using the correlation matrix. All 17 IVs were used in this run. Rows #1-7 are the 8-wk old R6/2 mice, whereas rows #8-13 are the 12-wk old R6/2 mice. The first principal component (Factor 1) correctly identified all of the 13 original R6/2 mice: all of the 8-wk old have positive factor scores, whereas all of the 12-wk old have negative factor scores. The rest of the factors (2-17) did not show any meaningful results with respect to the identification of the 13 original R6/2 mice.

Principal Components Report

Factor Score

	Factors					
Row	**Factor1**	**Factor2**	**Factor3**	**Factor4**	**Factor5**	**Factor6**
1	0.0691	0.0663	1.2098	-0.3138	-2.1068	0.2445
2	0.6811	-0.2991	-1.1961	0.8936	-1.2792	0.1691
3	1.1328	-0.1261	1.0001	0.1358	0.0253	-0.3761
4	0.8155	2.0385	-0.9072	1.2330	-0.4844	-0.1056
5	1.1035	0.8124	0.1207	-0.8665	0.9084	-0.5730
6	0.6106	0.2908	0.9260	-0.3148	1.4657	-0.0742
7	0.9830	-0.9506	0.2926	-1.0685	0.2363	1.2972
8	-0.1019	-0.7789	-1.7321	0.2884	1.1684	0.6819
9	-0.7522	-0.4548	-1.2823	-0.9016	-0.0508	-1.8014
10	-1.1369	-0.3866	1.2821	2.1115	0.8458	-0.8831
11	-0.6449	-0.9093	0.1597	-0.8343	-0.7399	-1.2457
12	-0.6969	-1.0750	0.0237	0.6719	-0.1157	1.6034
13	-2.0630	1.7724	0.1029	-1.0347	0.1269	1.0629

Table S9. Classification of the 30 original mice (17 WT & 13 R6/2) by the four principal components (factors) of ROC-supervised PCA using the correlation matrix. The top 4 IVs (Cr+PCr, Gln, Cr, and NAA) (AUC > 98%) according to the ROC curve analysis were used in this run. Rows #1-17 are the WT mice, rows #18-30 are the R6/2 mice. The first principal component (Factor 1) correctly identified all of the 30 mice: all WT mice have positive factor scores, whereas all R6/2 mice have negative factor scores. The rest of the factors (2-4) did not show any meaningful results with respect to the identification of the 30 original mice.

Principal Components Report

Factor Score

	Factors			
Row	**Factor1**	**Factor2**	**Factor3**	**Factor4**
1	1.0249	0.1820	1.2699	1.2244
2	0.1881	-0.1271	-0.6627	1.7080
3	0.5114	-0.4714	-0.3592	0.7711
4	0.7932	-0.4889	1.2512	1.2149
5	1.0988	-1.0721	0.3042	-0.3713
6	1.0520	-0.4155	-0.0120	-1.1467
7	0.9391	0.1321	0.2016	1.4978
8	0.8451	-0.2960	-0.6837	0.2755
9	0.9913	0.6241	0.4585	-1.1777
10	0.9107	-1.1555	-1.5490	-0.8547
11	0.8007	0.7661	0.2805	0.1813
12	0.7390	-0.4142	0.0684	1.2736
13	0.5222	0.7604	-0.1146	0.2500
14	0.5166	0.6072	-1.0581	0.3261
15	1.1343	1.0398	1.1330	-1.8964
16	1.0371	0.2848	-0.3989	-1.1634
17	0.9356	-0.4936	-0.4817	-0.4435
18	-0.4615	0.0298	0.6946	-1.0192
19	-1.1889	0.3436	-1.7104	0.3240
20	-0.7796	1.6990	0.1537	0.5984
21	-1.2258	1.8657	-0.2161	0.4688
22	-1.0471	0.9370	-1.0238	-0.2190
23	-0.9465	1.2680	1.0649	0.3214
24	-0.7114	0.6035	-0.3152	-1.0111
25	-1.4695	0.2652	-1.0467	-0.7360
26	-1.1024	-1.8187	-1.4806	1.1325
27	-1.5320	-0.6543	2.8068	1.1052
28	-0.9129	-1.8457	-0.3268	0.0697
29	-1.2994	0.2174	0.3348	-1.1889
30	-1.3632	-2.3728	1.4171	-1.5145

Table S10. Classification of the 13 original R6/2 mice [(7) 8-wk old & (6) 12-wk old] by the two principal components (factors) of ROC-supervised PCA using the correlation matrix. The top two most significant IVs (TTau and GPC+PC) (AUC > 95%) according to the R6/2 ROC curve analysis were used in this run. Rows #1-7 are the 8-wk old R6/2 mice, whereas rows #8-13 are the 12-wk old R6/2 mice. The first principal component (Factor 1) correctly identified all of the 13 original R6/2 mice: all of the 8-wk old have negative factor scores, whereas all of the 12-wk old have positive factor scores. The second factor did not show any meaningful results with respect to the identification of the 13 original R6/2 mice.

Principal Components Report

Factor Score

	Factors	
Row	**Factor1**	**Factor2**
1	-0.0946	-1.7039
2	-0.6453	1.1101
3	-0.8299	0.3600
4	-0.6423	1.7880
5	-0.9138	-0.3491
6	-0.5742	-0.9329
7	-1.5229	0.2086
8	0.2455	-0.0660
9	0.9764	0.9499
10	0.7147	-0.2327
11	0.3630	-1.2892
12	0.7339	-0.5796
13	2.1896	0.7370

Supplementary figures

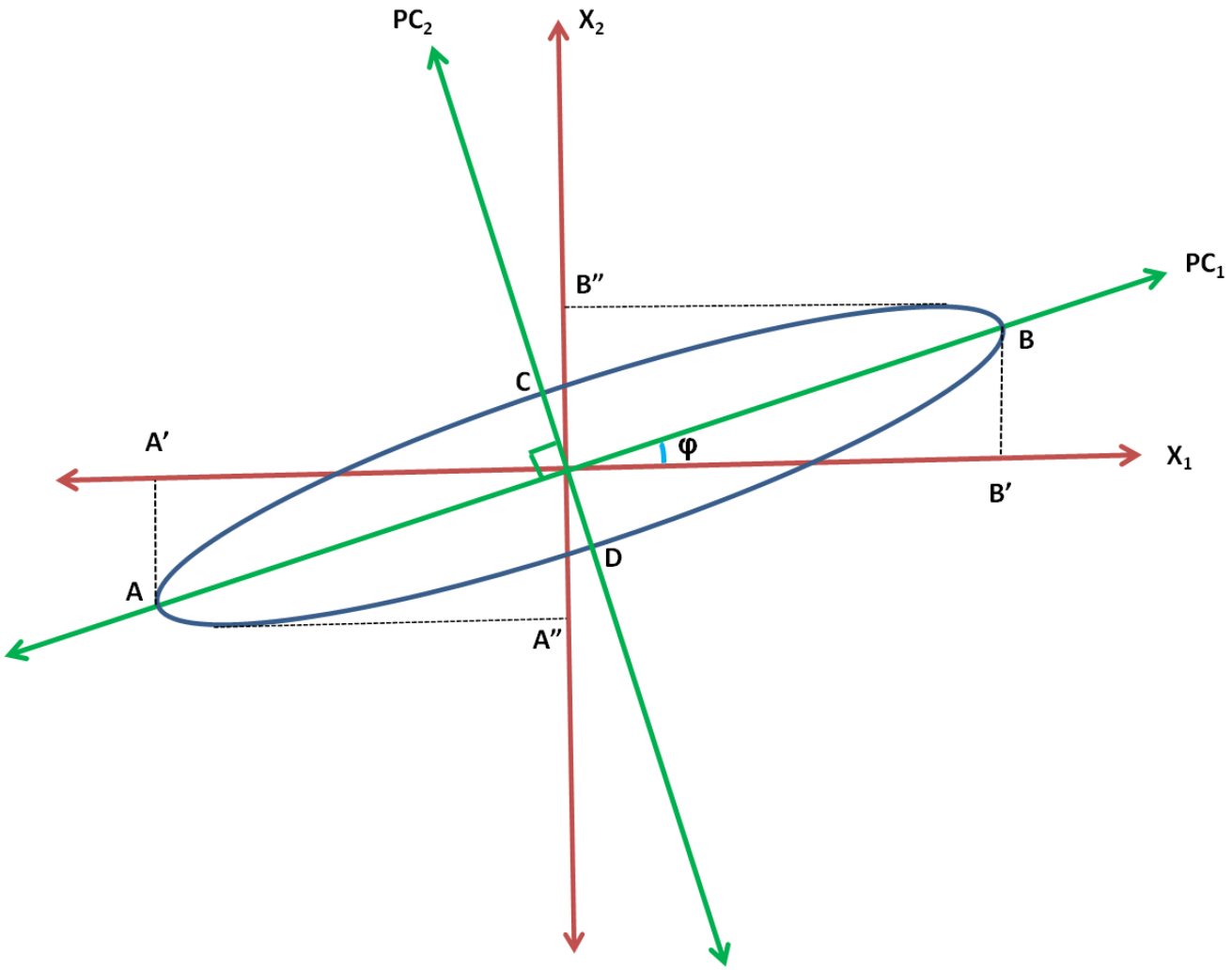

Figure S1. Geometrical representation of the formation of the principal components of PCA. Since variable X_1 has a larger variance than that of variable X_2 (A'B' > A"B"), PCA will seek to maximize the variance of X_1 by fitting the best straight line through the data points (contained within the blue-lined ellipse). That best straight line fitted is the first principal component (PC_1), whose variance is larger than that of the original variable X_1 (AB > A'B'). The second best fitted line through the data points with the proviso that it has to be perpendicular to PC_1 is the second principal component (PC_2).

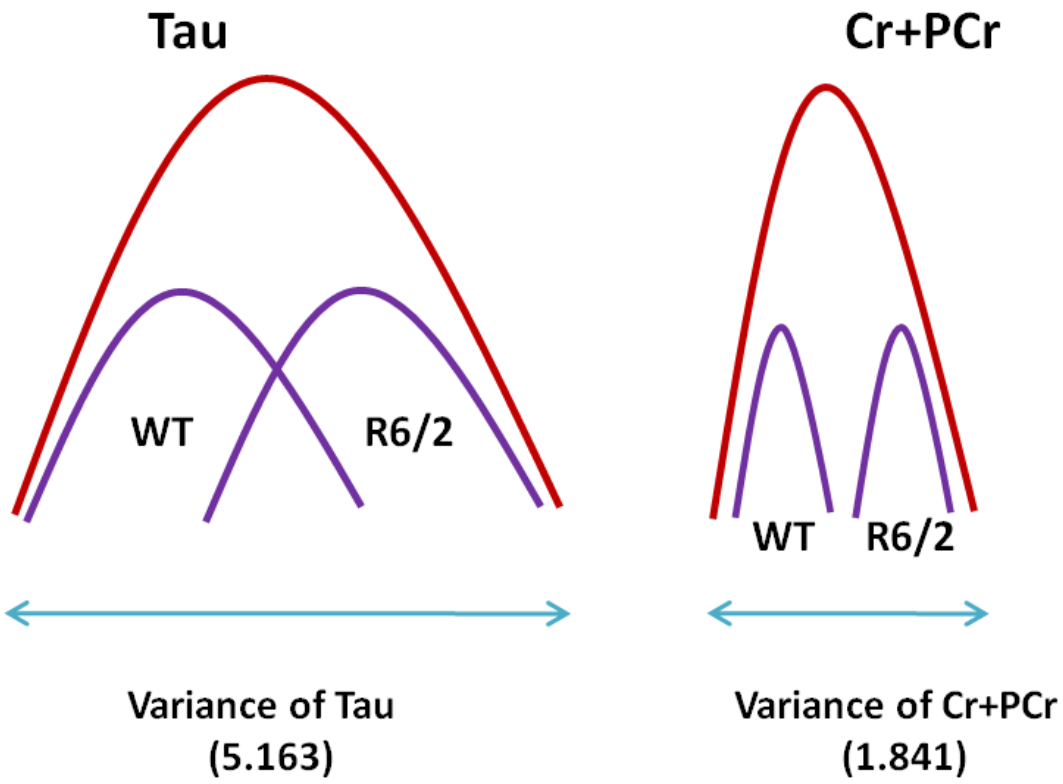

Figure S2. Diagrammatical representation of the Tau and Cr+PCr variables. There is a significant overlap between the two groups (WT & R6/2) in the Tau variable. Its AUC = 0.7289. In the Cr+PCr variable, on the other hand, there is no overlap between the two groups (WT & R6/2). The AUC of Cr+PCr is perfect (1.0000). For diagnostic purposes, therefore, Cr+PCr is the ideal biomarker, whereas Tau is not a significant biomarker. Because, however, Tau has a much larger variance than that of Cr+PCr, PCA will create the first principal component – the most significant of all principal components – in such a way that is influenced predominantly by the Tau variable.