GUIDE

Interpreting Breath Biopsy® Study Data

BREATH® BIOPSY

Across Breath Biopsy studies and the breath research field more broadly, there are a number of common graph types that are used to visualize and interpret results.

This short guide provides definitions for common statistical terms and readouts and provides additional commentary to allow successful and robust interpretation of results. It is intended as an introduction to the kind of data we report and key comparisons that are made within breath research.

Breath Biopsy studies include specialists from a wide range of backgrounds so we have developed this guide to be accessible to those that do not have existing experience of statistical analysis or reporting.

Terms and Definitions

Volatile Organic Compounds and Molecular Features

Breath Biopsy aims to discover, identify and validate volatile organic compounds (VOCs) as biomarkers on breath for use in early disease detection and precision medicine. The term VOC describes a wide range of chemicals, many of which can be found on breath.

Molecular Features (MFs) are peaks in the graphs produced by mass spectrometry analysis. Each peak is likely to indicate the presence of a specific chemical. Molecular features in each graph are referred to by number e.g. MF310.

In breath research, each MF can be expected to result from the presence of VOCs in the breath samples. Typically we refer to MFs in the early stages of analysis and only use VOCs once particular molecules have been identified and confirmed against a chemical standard.

Breath Biopsy studies typically either use Owlstone Medical's Breath Biopsy high resolution accurate mass (HRAM) Library of VOCs, or a library maintained by the National Institute of Standards and Technology, often referred to as the NIST Library to identify VOCs.

Adjusted and Unadjusted *P*-values

Many statistical tests seek to evaluate the difference between two groups of data. This is represented as the *P*-value, which is essentially the probability of you getting your results assuming that the null hypothesis is true i.e. there is no difference between the groups. As such a small *P*-value suggests that there is more likely to be a difference between the two groups.

When the same statistical test is conducted repeatedly for many MFs, the probability of falsely finding a difference (false discovery) increases. To compensate for this, we can use adjusted *P*-values. This adjustment for 'multiple testing' corrects the effect of using the same statistical test many times.

Although adjusted *P*-values provide higher confidence, they can be overly restrictive, especially in studies with small numbers of samples. As such, early stage biomarker discovery studies often use unadjusted *P*-values. This reduces the risk of excluding VOCs that could turn out to be of interest when investigated in further studies with a larger number of samples.

Using *P*-values

In many cases from published literature, any test that results in a *P*-value <0.05 is interpreted as having a significant association. However, a *P*-value of 0.05 for any single statistical test means there is still a one in 20 chance that there is no biological difference between the groups being compared.

We seek to avoid overinterpreting *P*-values and, where possible, try to provide biological evidence or rationale to support results. This is often achieved by identifying multiple VOCs linked to the same processes or pathways.

Principal Component Plots

Definition: Principal component analysis (PCA) summarises complicated data by representing the data using fewer dimensions, a process called unsupervised dimensionality reduction. PCA can provide a visual summary of variation within a dataset.

The data from each MF in a study can be represented as a dimension on a graph, one dimension for each MF. Other factors such as age or sex are also dimensions. PCA reduces the number of dimensions by combining groups of correlated MFs into principal components (PCs).

The first principal component (PC1) accounts for as much of the variation in the data as possible. The subsequent PCs account for as much of the remaining variation, with some restrictions. As the majority of variation will be limited to the first few PCs, plots often only include PC1 and PC2 (**Fig. 1a/1b**).

Interpretation: In Breath Biopsy, PCA plots are used to visually inspect the underlying structure of the data. The presence of structure can indicate limitations in the study design, sample handling or chemical analysis that will need to be considered in the statistical analysis. Ideally there should be no structure in the data as this means that there are no broad differences between the groups of samples.

Each breath sample is represented as a point on the plot. Structure can appear as patterns of related points, which can include separate clusters of points related to different sample categories (**Fig. 1c**).

PCA can only be used to see large differences between samples. The differences caused by changes to individual MFs are too subtle to be identified using this approach.

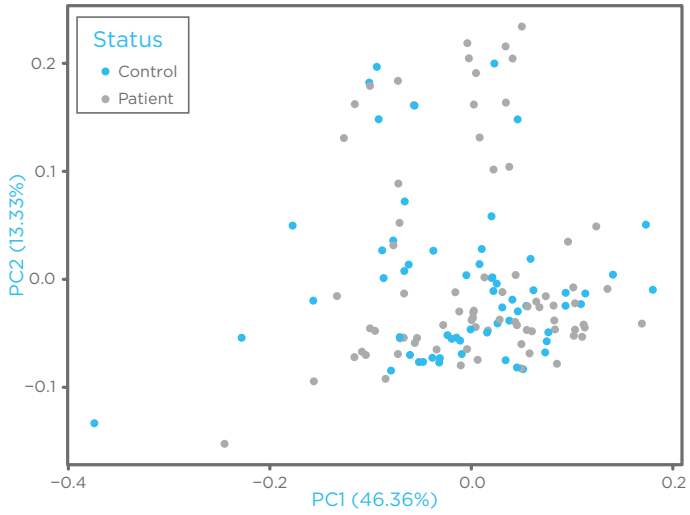


Figure 1a. A principal component plot for breath case-control study analysis. The percentage of the total variation in the data that is represented by each principal component is shown in the titles of the axes. The data shown here do not show structure for case vs. control as there are no distinct clusters of points. The same plot could be coloured according to other characteristics to see if any other structures can be identified. Outliers are excluded from this plot.

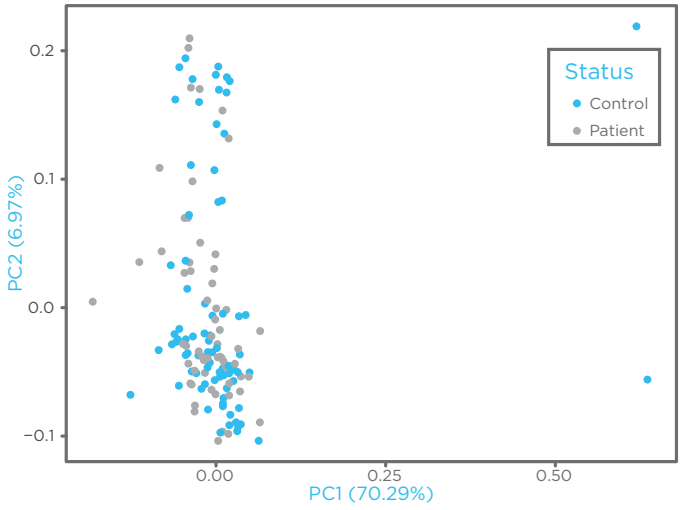


Figure 1b. A principal component plot for breath case-control study analysis. This plot shows the same data as Figure 1a but with outliers included (two points on the right of the plot).

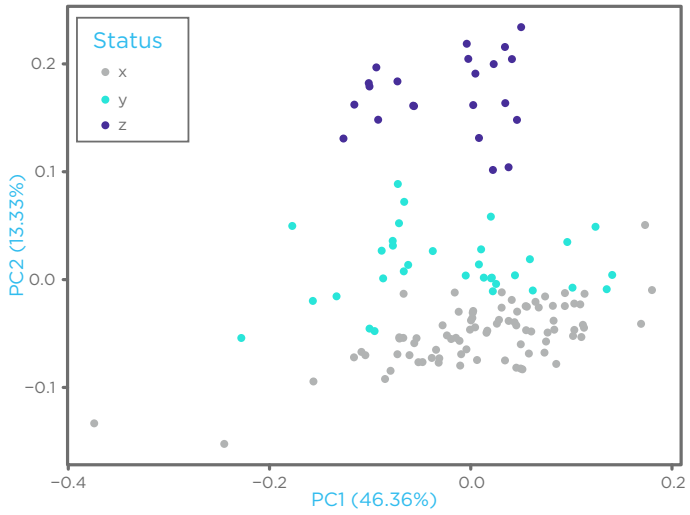


Figure 1c. A principal component plot for breath case-control study analysis showing structure. The same plot as Figure 1a but coloured by a different set of statuses (x,y and z). This plot shows strong structure with all samples from each group occurring together on the plot.

The difference between the groups could be biological (e.g. age group, sex, race, smoking status) or could be an artefact of the study (e.g. sample collection time, or analysis sequence).

Where differences like this are found, it's important to maintain awareness of them during the rest of data analysis to try and consider how this might affect results.

Volcano Plots

Definition: Volcano plots provide a visual summary of a test for differences between cases and controls. The plots relate the magnitude of a change (fold-change) and the statistical significance (*P*-values) of the change for each molecular feature (MF) tested (**Fig. 2a**). An alternative version of this plot for continuous variables (e.g. FEV1 lung function scores), uses a standardized regression coefficient in place of fold-change, but the interpretation is broadly similar (**Fig. 2b**).

Interpretation: Each point represents an MF. The *P*-values and fold-changes, displayed on the y and x axes respectively, are usually presented using logarithms to make the plots easier to interpret. Thresholds for *P*-value and fold-change are typically applied to identify MFs of interest. The *P*-value threshold will usually be based on an adjusted *P*-values. The MFs of interest are the points beyond the thresholds (shown in orange below).

Numbers:

- A $\log_2$(fold-change) of 1 on the x-axis occurs when a particular MF has twice the abundance in cases compared to that in controls.
- Negative $\log_2$(fold-change) changes occur where an MF is reduced in cases compared to controls. A value of -1 means the MF has half the abundance in cases relative to controls.
- A -$\log_{10}$(*P*-values) of 1.30 on the y-axis equates to a *P*-values of 0.05 and a value of 2 equates to a *P*-values of 0.01.

Illustrative Plots:

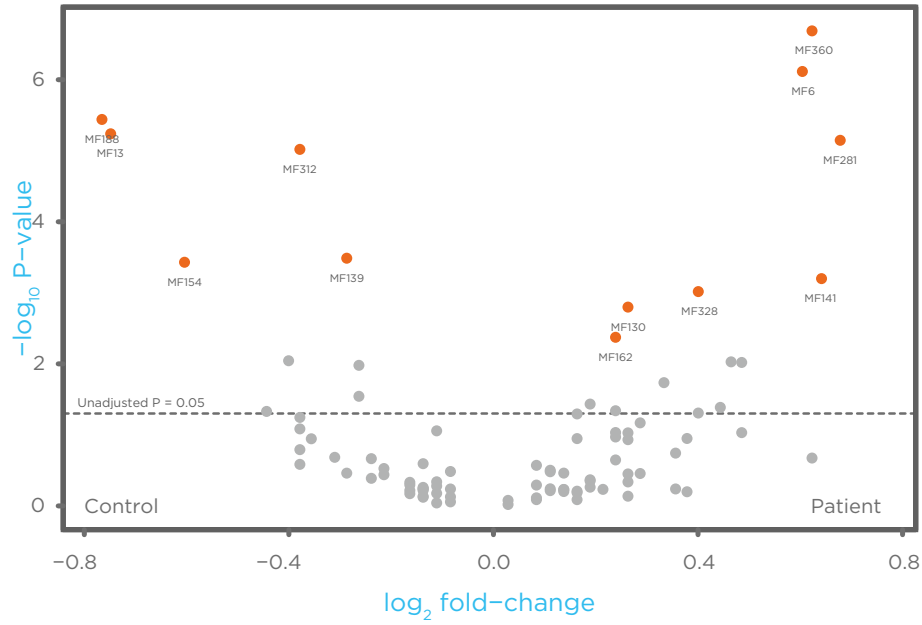


Figure 2a. A volcano plot for a breath case-control study analysis. The dashed horizontal line indicates an unadjusted *P*-value of 0.05. MFs considered to be significant based on the adjusted *P*-value are shown in orange.

The fold change for each MF is calculated as the difference between the average abundance of the MF in the control group vs. the case group.

The labels Control and Patient on plot indicate which group of samples has the higher abundance for MFs on that side of the plot.

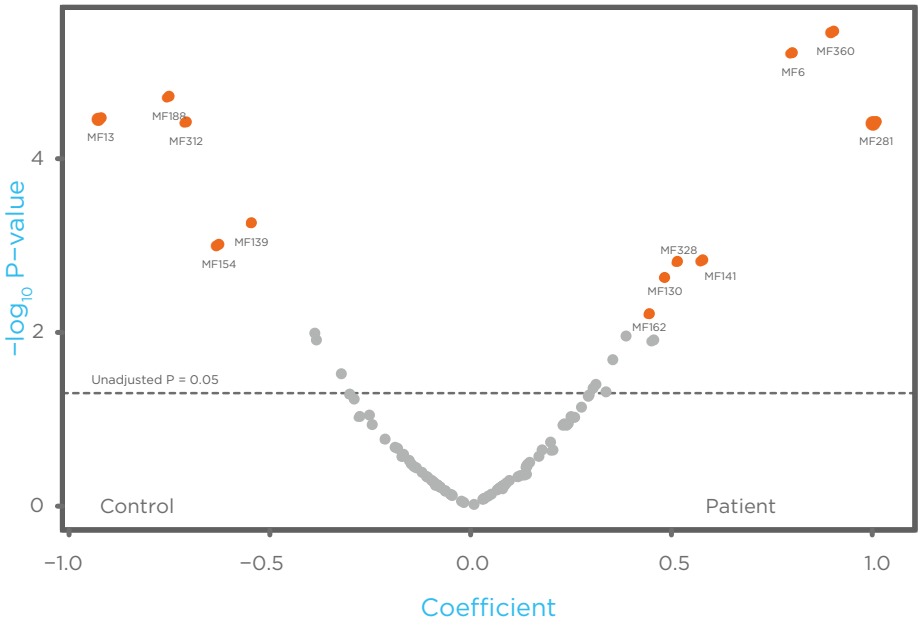


Figure 2b. A volcano plot for a continuous variable.

Continuous variables are factors, such as height, which can have many different values. Because fold-change is calculated by dividing the average of one group (cases) by the average of another (controls), it is no longer relevant for continuous variables.

Instead of fold-change, volcano plots of continuous variables use a standardized regression coefficient. Similar to fold-change, a larger coefficient corresponds to a greater variation in the abundance of an MF.

For example, FEV1 measures the amount of air that can be expelled from the lungs in a second. The result (voulme of air) is a continuous variable. MFs that vary a lot between patients with low FEV1 and those with high, will have a large coefficient. While MFs that vary less will have a smaller one.

Volcano plots using standardized coefficients often have a more distinct 'V'-shape compared to those based on fold-change. MFs considered to be significant based on the adjusted *P*-value are shown in orange.

Box Plots

Definition: Box plots summarise the distribution of a continuous variable, such as the abundance of a VOC.

The first and third quartiles define a box in the middle of the plot, which represents the distribution of the middle 50% of the data in each group. The median is a horizontal line that crosses the box. Half of the values are greater than or equal to the median and half are less. The other components of the plot are defined in **Figure 3a**.

Interpretation: Each plot represents the abundance of an MF in a particular group of samples. Plots that span a large range vary more between samples. These plots can be used to visualise how individual MFs differ between groups of samples. Example plots are shown in **Figure 3b**. The MF shown in these plots has a moderately higher abundance in cases relative to controls.

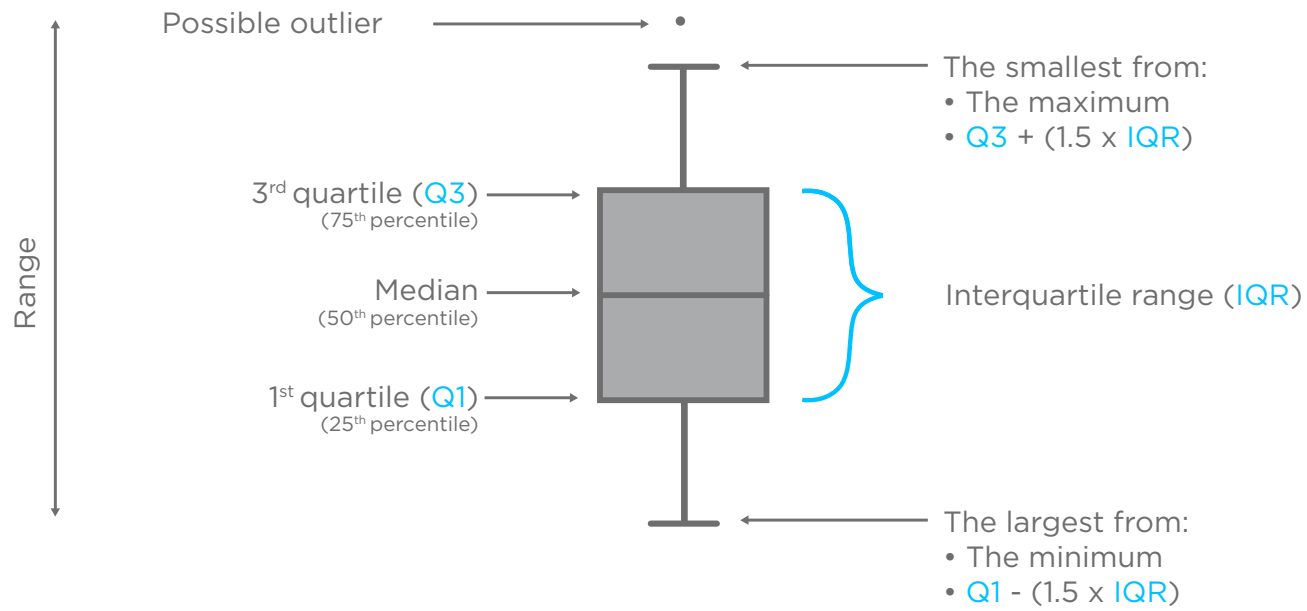


Figure 3a. The components of a box plot defined.

Illustrative Plots:

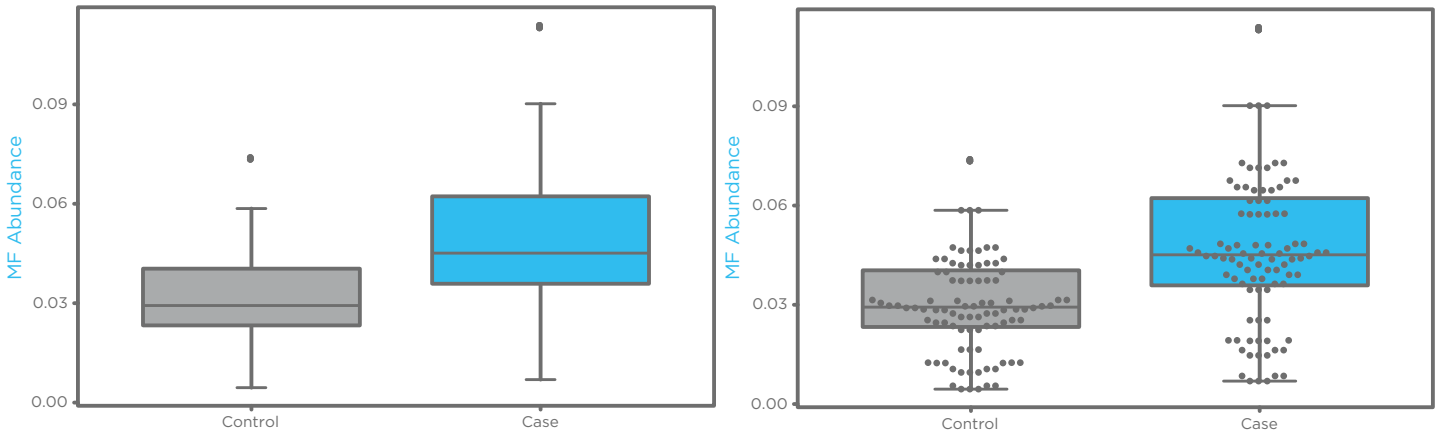


Figure 3b. Box plots for an MF from a breath case-control study analysis. In the right-hand panel, the individual MF abundance measurements are shown in full as grey points. In this example, the MF abundance in the case population (cyan) is typically higher than in the control population (grey).

Receiver Operating Characteristic (ROC) Curves

Definition: The ROC curve can be used to assess the results of binary classifiers, which are computerized tools for splitting data into two groups, such as distinguishing cancer samples from non-cancer. The plot displays sensitivity (also known as the true positive rate or recall) and 1 - specificity (also known as the false positive rate) (**Fig 4**).

Sensitivity reflects the proportion of positive samples that are correctly identified. Specificity reflects the proportion of negative samples that are correctly identified, as such 1-specificity is the number of negative samples incorrectly identified as positive. An ideal test should have high sensitivity and specificity.

Any binary classifier needs a threshold to judge positive results from negative, when the threshold is low all samples are reported as positive, when it is high all samples are counted as negative. Each point along an ROC curve shows the sensitivity and specificity that result from using the classifier with a given threshold value.

Interpretation: Binary classifiers try to tell the difference between case and control samples based on the available data. Classifications can be assigned correctly or incorrectly. The correct results are true negatives and true positives, while errors are false positives and false negatives.

Classifiers can be compared by plotting several models on the same axes. Classifiers with a high area under the curve (AUC) are generally better at making correct judgements.

Numbers:

- Both sensitivity and specificity can have values between 0 and 1 or can be reported as percentages.
- Like sensitivity and specificity, the accuracy of a model can also be evaluated at a specific threshold value. Accuracy reflects the proportion of all samples that are correctly assigned.
- A classifier with AUC=0.5 is equivalent to random guessing. Interpretation varies depending on the clinical requirements of a test, a lower AUC may be sufficient for a test that will be used for large scale screening in advance of more invasive and expensive investigations. However, AUC values can be broadly interpreted as:
  - 0.6-0.7: The model is poor at assigning positive or negative results
  - 0.7-0.8: The model is fair at assigning positive or negative results
  - 0.8-0.9: The model is good at assigning positive or negative results
  - 0.9-1.0: the model is excellent at assigning positive or negative results

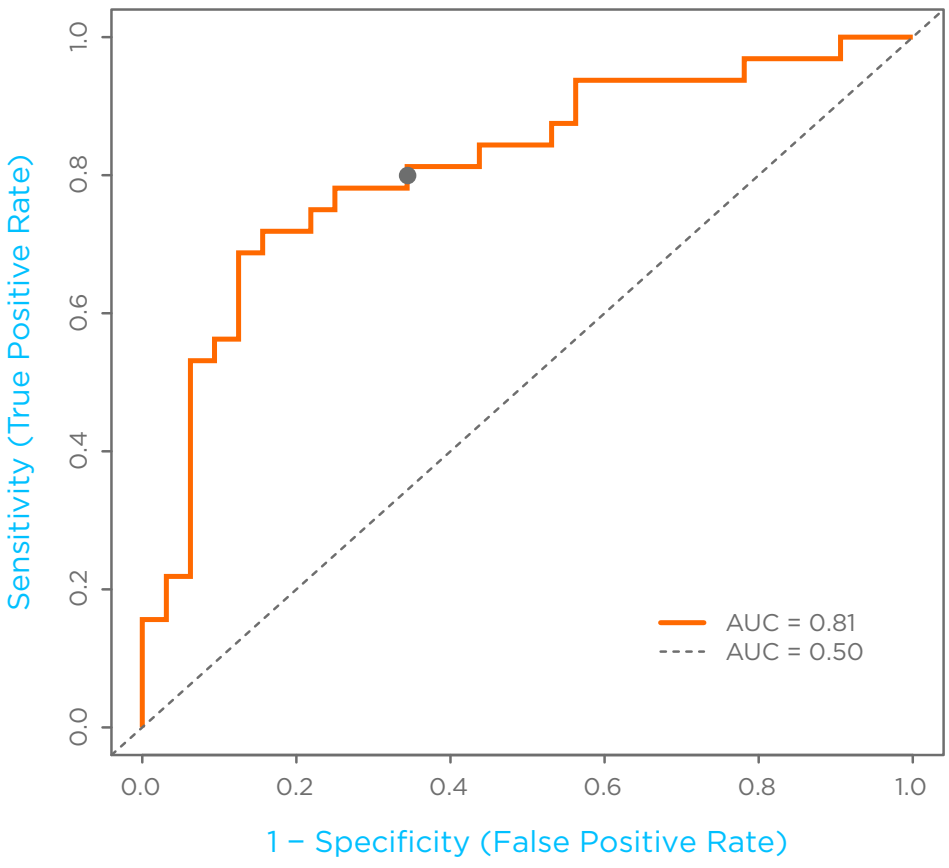


Figure 4. A receiver operating characteristic curve from breath case-control study analysis. The curve represents changes in true positive and false positive rates when using a binary classifier model and changing the threshold for assigning results. Each point on the graph is the result of using a different threshold.

At the point marked on the plot, sensitivity is 0.8 and 1-specificity is around 0.35. This means that 80% of positive samples will be correctly identified as positive, and only 35% of negative samples will be incorrectly identified as positive.

The grey diagonal line represents AUC=0.5, where the performance of the model is no better than random chance. The AUC for this classifier is 0.81 meaning it has a good ability to differentiate cases from controls.

Feature Tables

Feature	Retention Time	ID (Breath Biopsy Library)	CAS	Fold Change	P-value	Adjusted P-value
MF1	38.378	D-Limonene	5989-27-5	0.841	0.352	0.473
MF2	43.3	1-Hexadecanol	36653-82-4	1.523	0.010	0.222
MF3	36.095	2-Butanone	78-93-3	0.484	0.127	0.404
MF4	56.698	Isopropyl alcohol	67-63-0	1.522	0.037	0.295
MF5	10.715	Pentane, 3,3-dimethyl-	562-49-2	0.617	0.251	0.462
MF6	47.045	Isoprene	78-79-5	0.661	0.105	0.398
MF7	8.335	Heptane, 2-methyl-	592-27-8	0.837	0.015	0.042
MF8	18.617	Undecane	1120-21-4	0.589	0.168	0.404
MF9	51.96	p-Xylene	203-396-5	0.823	0.374	0.482
MF10	18.249	Cyclohexanone	108-94-1	0.887	0.010	0.067

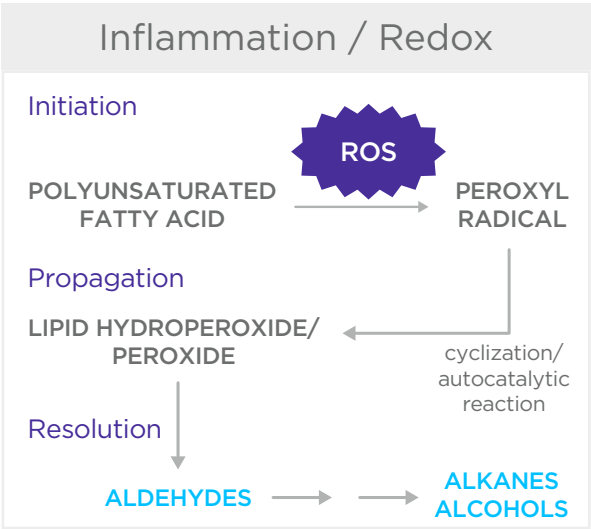
All Breath Biopsy project reports are accompanied by a full feature table covering all molecular features identified in the breath samples.

IDs are provided indicating the VOC most likely to be associated with each molecular feature. IDs are often assigned by comparing collected data to a record of known VOC standards. These standards can either be from the Breath Biopsy HRAM Library (as shown here) or the online National Institute of Standards and Tecnology (NIST) mass spectral library, in which case a measure of confidence is also included.

CAS numbers are a standardized system for identifying chemicals since many common compounds have several different names.

Our feature tables also include key values from our analysis which may include fold-changes, regression coefficients or *P*-values, as well as integrated peak areas from the GC-MS data. This means you have all the data available for your own analysis and to integrate with other results in your wider study.

Biological Interpretation



With our more advanced reporting options, we can provide more detailed expert commentary and biological interpretation. We can help to identify groups of key VOCs that may be worthy of further investigation.

Our interpretations can include identification of key VOCs that may be worthy of further investigation or summaries of observations in the scientific literature to suggest how changes represent disease-related changes in biological pathways (**Fig. 5**). These conclusions can help guide further studies and allow the design of more focused validation studies.

Figure 5. An example biological pathway graphic included with our interpretations. This example demonstrates the production of aldehydes, alkanes and alcohols from the lipid peroxidation pathway, a notable component of inflammation and oxidative stress, which are relevant in many disease contexts.

Owlstone Medical Ltd, 183 Cambridge Science Park,
Milton Road, Cambridge, CB4 0GJ, UK
Company Number 04955647 | VAT Number 260449214

BREATH® BIOPSY