

# Quantitative gene expression results using TITAN: A quick user guide

Gareth Elvidge and Tom Price, August 2004

These are instructions for running TITAN using **R** on a Windows platform.

## 1. Installing R and TITAN

- 1.1. Download and install the most recent version of **R** from <http://www.cran.r-project.org>
- 1.2. Download and save the latest compiled version of the R for Windows package **titan** from <http://www.well.ox.ac.uk/~tpice/titan>
- 1.3. Run **R**. Install the packages **VR**, **tcltk**, **boot** and **splines** from CRAN using the menu option  
**Packages > Install package(s) from CRAN...**  
from the R toolbar. Install the package **titan** from the saved zip file using the menu option  
**Packages > Install package(s) from local zip files...**
- 1.4. Load the library by typing  

```
library(titan)
```

at the R console.
- 1.5. Read the online help files on the program and on the example dataset by typing  

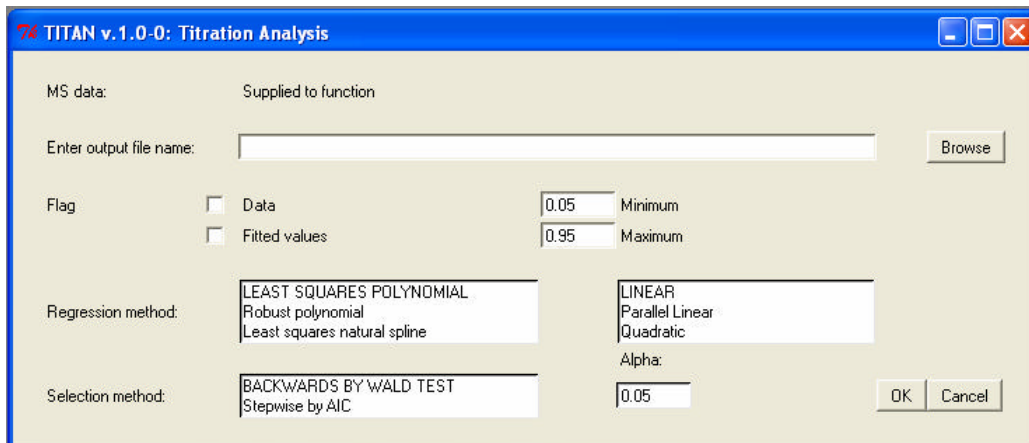
```
?titan
```

```
?hypoxia
```

## 2. Running an example analysis

- 2.1. Run the example from the online help file, which uses a sample data set, by typing  

```
example(hypoxia)
```
- 2.2. A GUI (graphical user interface) will appear and prompt you for options to use in the initial data analysis. The input box that appears will look something like the screen shot below.



- 2.3. R To save the graphical output as a pdf file, enter a file name in the entry box **Enter MS data file name**. For now, leave the output file blank: this ensures that the regression graphs will be displayed within **R**. If you like, you can change the data-flagging and regression parameters. These parameters affect the regression lines that are used to interpolate the equivalence points in the data.
- 2.4. When you have finished, press **OK** or hit **Return**. A new input box will appear (see below). The interface prompts you to select baseline and test treatment conditions, and test and housekeeping (control) genes. These parameters are important in calculating the fold changes between different conditions, adjusting for the results of the housekeeping genes. You will also be asked to supply parameters for the bootstrap analysis that determines the statistical significance and confidence intervals of these fold changes. For this example, it is better not to set the number of bootstrap replicates too high or the analysis will proceed very slowly.

**TITAN v. 1.0-0: Titration Analysis**

Baseline treatment:

Comparison treatments:  Normoxia  Hypoxia

	T	H		T	H
Select Test and Housekeeping genes:	<input type="checkbox"/>	<input checked="" type="checkbox"/>	BMP2	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>	BNIP3	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>	CA9	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>	EGLN1	<input type="checkbox"/>	<input checked="" type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>	EGLN2	<input type="checkbox"/>	<input checked="" type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	EGLN3	<input checked="" type="checkbox"/>	<input type="checkbox"/>
				<input checked="" type="checkbox"/>	<input type="checkbox"/>
				<input type="checkbox"/>	<input checked="" type="checkbox"/>
				<input checked="" type="checkbox"/>	<input type="checkbox"/>
				<input type="checkbox"/>	<input checked="" type="checkbox"/>
				<input checked="" type="checkbox"/>	<input type="checkbox"/>
				<input type="checkbox"/>	<input checked="" type="checkbox"/>
				<input checked="" type="checkbox"/>	<input type="checkbox"/>

Confidence intervals:  95%

Confidence interval types:  norm  
 basic  
 perc  
 bca

Number of bootstrap replicates:

Seed for random number generator:

- 2.5. After a short delay a dialog box will appear, asking if the user wants to review the flags. By selecting **Yes** the user can cycle through the regression curves for each gene and select/deselect data points to be included in the analysis by pointing and clicking Data points

represented by crosses are excluded from the analysis and data points represented by circles are included in the analysis. The new flags can then be saved to the data file.

- 2.6. Next, the numerical output is printed in the **R** console and the regression graphs are output either to a graphical window in **R** or as a pdf file, depending on the previous selection.
- 2.7. The initial data analysis is now complete. The example goes on to rerun the analysis two more times without prompting you further. The first reanalysis is simply a demonstration of the syntax that enables you to repeat an analysis without requiring user input, and gives identical results to the initial analysis. Finally, a second reanalysis is performed that alters two of the parameters before performing the data analysis a third time.

### 3. Running an analysis on your own data

- 3.1. Users wishing to analyse their own competitive-PCR data (independent of detection technique used) need to prepare the data file in the following format. Create a tab-delimited text file containing the following column headers, with the sample-related information following in the rows below. The file may also contain other columns, but unless the column headings have similar names to the columns listed below, they will simply be ignored.

**FREQUENCY**: Relative frequency of the test cDNA and competitor concentration peaks *e.g.* 0.4

**GENE**: gene name *e.g.* GAPDH

**TREATMENT** or **RX**: sample type/name *e.g.* hypoxia/normoxiaC012

**COMPETITOR** or **CONCENTRATION**: concentration of competitor in moles *e.g.* 1E-15

**FLAG**: optional data quality flag. A value of 1 excludes the line of data from the analysis, whereas a value of 0 includes the data.

*Note*: if you are using the MassARRAY<sup>TM</sup> system to analyse samples, most of this information is given automatically by 'Genotype Analyser' (MassARRAY<sup>TM</sup>, Sequenom<sup>TM</sup>). Extra columns will need to be created to include the COMPETITOR / CONCENTRATION and TREATMENT / RX columns. This can easily be done in Microsoft Excel, and the resulting file saved as a text file.

- 3.2. Now analyse your data by typing something like

```
t = titan()
```

at the R console. You can choose any variable name you like in place of *t*.

- 3.3. The analysis using the GUI will proceed much as previously, except that the first input box will prompt you to enter the location of your data file in the entry box **Enter MS data file name**.

### 4. A quick guide to data interpretation

TITAN transforms the cDNA frequencies  $f$  according to the formula:

$$y = \frac{\log_{10}(f)}{\log_{10}(1-f)}$$

The transformed values are entered into a linear regression equation as the dependent variable, with the logarithm of the competitor concentration as the independent variable. The equivalence point is given by the concentration at which  $y = 0$  and is calculated by interpolation. If multiple conditions or genes are entered in the analysis, then multiple terms are entered into the regression to estimate separate slopes for each combination of gene and treatment condition.

TITAN outputs the following statistics for each gene:

Regression results per gene:				
	Sum Sq	Lack of Fit	Residuals	R-Squared
BMP2	34.79	0.004604	0.012996	0.9824
BNIP3	43.89	0.007770	0.008077	0.9842
CA9	38.69	0.005851	0.011655	0.9825
EGLN1	20.13	0.014090	0.005464	0.9804
EGLN2	31.89	0.013495	0.003054	0.9835
EGLN3	40.45	0.009411	0.005339	0.9853
HFE	13.95	0.002830	0.012965	0.9842
HIF1A	19.57	0.003040	0.007433	0.9895
NDRG1	36.78	0.005911	0.008434	0.9857
PPPCC	47.78	0.004996	0.011831	0.9832
SLC3A2	42.09	0.009676	0.005440	0.9849
VEGF	31.13	0.009959	0.005538	0.9845
Total	401.16	0.007714	0.008187	0.9841

Sum Sq Total sum of squares for the transformed data

Lack of fit Proportion of variability due to deviation of data groups from the regression line

Residuals Proportion of variability due to deviation between replicate reactions

R-squared Proportion of the variability that can be explained by the regression line

R-squared, residuals and lack-of-fit are all expressed as a proportion of the sum of squares of the gene-specific data and sum to 1 before rounding. A good quality assay should have good concordance between replicate PCR reactions (i.e. a small residual value, less than ~0.05 perhaps) and the means of replicate samples should lie close to the regression line (i.e. a lack-of-fit value less than ~0.05).

The output for the interpolation is as follows:

Interpolated Concentrations:		
	Normoxia	Hypoxia
BMP2	4.553227e-17	3.380852e-17
BNIP3	1.523064e-15	1.102839e-14
CA9	2.641116e-16	1.700027e-14
EGLN1	5.378075e-15	1.768936e-14
EGLN2	2.063165e-15	1.993495e-15
EGLN3	4.401633e-15	2.841423e-14
HFE	2.781709e-17	4.964441e-17
HIF1A	4.173086e-15	2.003979e-15
NDRG1	7.183796e-16	3.789322e-14
PPPCC	6.988261e-15	4.990734e-15
SLC3A2	5.545915e-15	1.035343e-14
VEGF	7.798583e-14	2.193162e-13

Fold Change:	
	Hypoxia
BMP2	0.7425177
BNIP3	7.2409200
CA9	64.3677626
EGLN1	3.2891626
EGLN2	0.9662314
EGLN3	6.4553850
HFE	1.7846727
HIF1A	0.4802152
NDRG1	52.7481776
PPPCC	0.7141596
SLC3A2	1.8668569
VEGF	2.8122564

Interpolated concentrations refers to the equivalence points, calculated by interpolation. Fold changes are calculated relative to the baseline condition, adjusted for the fold changes of the housekeeping genes.

The bootstrap routine resamples the residuals from the regression at random to simulate new datasets, called 'bootstrap samples'. When the regression is re-run on these simulated datasets, the fold changes

that result have a spread of values from which we can determine the statistical significance of the original fold changes. TITAN's bootstrap routine gives the following output:

Bootstrap statistics:						
	Fold	Log10 Fold	Bias	Std.Error	p	
Gene BMP2, Rx Normoxia	0.7425177	-0.12929317	-0.0017228487	0.04527028	0.000	
Gene BNIP3, Rx Normoxia	7.2409200	0.85979375	0.0008821078	0.03607798	0.000	
Gene CA9, Rx Normoxia	64.3677626	1.80866841	0.0014165881	0.04201390	0.000	
Gene EGLN1, Rx Normoxia	3.2891626	0.51708534	-0.0017185578	0.04534436	0.000	
Gene EGLN2, Rx Normoxia	0.9662314	-0.01491887	0.0006651788	0.03379973	0.336	
Gene EGLN3, Rx Normoxia	6.4553850	0.80992215	-0.0023820561	0.03510843	0.000	
Gene HFE, Rx Normoxia	1.7846727	0.25155859	0.0001689786	0.05402657	0.000	
Gene HIF1A, Rx Normoxia	0.4802152	-0.31856413	0.0001539626	0.04094340	0.000	
Gene NDRG1, Rx Normoxia	52.7481776	1.72220746	0.0014052748	0.03385135	0.000	
Gene PPPCC, Rx Normoxia	0.7141596	-0.14620473	0.0009318522	0.02858967	0.000	
Gene SLC3A2, Rx Normoxia	1.8668569	0.27111103	-0.0001199159	0.03060939	0.000	
Gene VEGF, Rx Normoxia	2.8122564	0.44905491	0.0016904425	0.03772314	0.000	

- Fold** Fold change as determined by the ratio of the interpolated concentrations.
- Log 10 Fold** Base 10 logarithm of the fold change.
- Bias** The difference between the mean of the log10 fold changes calculated from the bootstrap samples, and the original log10 fold change. If the bootstrap routine has been performed successfully, this should take a value close to zero.
- Std.Error** Standard error of the bootstrap log10 fold changes.
- p** Bootstrap p-value. A significant fold change is one for which this value is near zero: typically a threshold of  $p < .05$  is used to determine statistical significance. The accuracy of this calculation is limited by the number of bootstrap replicates.

Confidence intervals can also be calculated based on the bootstrap routine. There are several different methods of calculating the fold changes: the 'basic' confidence intervals probably give results that are most consistent with the bootstrap p-values. Note that the confidence intervals can also be different when calculated on fold changes than when calculated on the log fold changes.

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS							
Fold change:							
Level	Normal			Basic			
95%	(	0.6077173,	0.9144454	)	(	0.6060706, 0.9180392	)
Level	Percentile			BCa			
95%	(	0.6005545,	0.9096838	)	(	0.6061357, 0.9144321	)
Log10 fold change:							
Level	Normal			Basic			
95%	(	-0.2162984,	-0.0388422	)	(	-0.2174768, -0.0371388	)
Level	Percentile			BCa			
95%	(	-0.2214476,	-0.0411096	)	(	-0.2174301, -0.0388485	)