Florida Atlantic Harbor
S. Edge
5600 N. US Hwy 1
FL 34946 FORT PIERCE
VER. ST. VAN AMERIKA

Project code

54114

Date of Report

07-feb-2013

Ordernumber Client

Next generation sequencing project report

Authorization

Marten Boetzer
Bioinformatician

Walter Pirovano
Productspecialist

BaseClear Group
Einsteinweg 5
2333 CC Leiden
The Netherlands
T +31 71 523 3917
F +31 71 523 5594
E info@baseclear.com

1. Introduction

This report contains a detailed overview of your Next generation sequencing experiment. It includes the following analysis (specified per sample).

	Analysis 1	Analysis 2
buc	Quality analysis of FASTQ sequence reads	RNA-Seq analysis
bod	Quality analysis of FASTQ sequence reads	RNA-Seq analysis
hoc	Quality analysis of FASTQ sequence reads	RNA-Seq analysis
35ppt	Quality analysis of FASTQ sequence reads	RNA-Seq analysis
30ppt	Quality analysis of FASTQ sequence reads	RNA-Seq analysis
25ppt	Quality analysis of FASTQ sequence reads	RNA-Seq analysis

For each analysis a short summary is given in the following section(s). Also detailed analysis statistics are provided per sample as Enclosure.

2. Quality analysis of FASTQ sequence reads

The FASTQ sequence reads were generated using the Illumina Casava pipeline version 1.8.2. Initial quality assessment was based on data passing the Illumina Chastity filtering. Subsequently, reads containing adapters and/or PhiX control signal were removed using an in-house filtering protocol. The second quality assessment was based on the remaining reads using the FASTQC quality control tool version 0.10.0. The final quality scores per sample are provided as Enclosure.

3. Transcriptome assembly

De novo transcriptome assembly was performed using Trinity (r2011-08-20). For each group a consensus assembly was created which represents the consensus transcriptome. Annotation of the consensus transcriptome was performed using BLASTall (blast-2.2.24) against the UniProt/SwissProt database.

4. RNA-Seq Analysis

The quality of the FASTQ sequences was enhanced by trimming off low-quality bases using the “Trim sequences” option of the CLC Genomics Workbench version 5.5.1. The quality-filtered sequence reads are used for further analysis with the CLC Genomics Workbench. First an alignment against the reference(s) and calculation of the expression values has been performed using the “RNA-Seq” option. Subsequent comparison of expression values and statistical analysis have been performed with the “Expression analysis” option.

The selected expression measure is the RPMK. It is defined as the Reads per Kilobase of exon model per Million mapped reads (Mortazavi *et al.*, 2008) and seeks to normalize for the difference in number of mapped reads between samples as well as the transcript length. It is given by dividing the total number of exon reads by the number of mapped reads (in Millions) times the exon length (in kilobases).

Quality control was performed to examine the consistency of the experiment and the variability between samples and groups. In Brief, the overall distribution of the RPMK expression values is compared between samples and groups through a boxplot representation. Also a principle component analysis is performed using the RPMK values. Finally samples are clustered into groups using a hierarchical clustering approach.

In the Enclosure the following sheets are provided:

- An explanation of the quality control figures of the RNA-Seq experiment. A number of analyses are performed to compare the RPMK expression values between samples (and groups if applicable). The analysis include:
    - Comparison of the overall distribution of RPMK expression values
    - Principle Component Analysis
    - Hierarchical sample clustering
- An explanation of the headers in the RNA-Seq analysis table. Here you can find more details about the expression and statistical measures used.
- The RNA-Seq alignment statistics per sample.

5. Results

A Summary of the results is provided below. Sequence and project data are recorded digitally in our secure database and stored for backup purposes only. Data is stored for a period of one year. The result files which have been generated within this project are (for each analysis) as followed:

<u>Raw sequence reads</u>

The files are stored in the "raw_sequences" folder and include:

- The Raw sequence reads in FASTQ format (*.fastq)

<u>Transcriptome assembly</u>

The files are stored in the "transcriptome_assembly" folder and include:

- The contig sequences in FASTA format (*_contig-sequences_Trinity.fa)
- The annotation files for the two groups in GFF format (*_contig-sequences_Trinity.gff)

<u>RNA-Seq analysis</u>

The alignment files are stored in the "reference_alignment" folder and include per sample:

- The alignment(s) in BAM format (*alignment_bam.bam)

The RNA-Seq analysis files are stored in the "RNA-Seq" folder and include:

- The full RNA-Seq analysis table containing the (normalized) expression values <u>between</u> the samples in Excel format (*rnaseq-comparison_final.xlsx)
- <u>Per sample</u> the individual expression tables in Excel format (*rnaseq-table.xlsx)

The quality control analysis figures are stored in the "analysis_figures" folder and include:

- The boxplot figure showing the comparison between the original RPMK values in PNG format (boxplot_RPMK-original.png)
- The boxplot figure showing the comparison between the root-mean square transformed RPMK values in PNG format (boxplot_RPMK-transformed.png)
- The Principle Component Analysis (PCA) plot in PNG format (PCA_RPMK-transformed.png)
- The hierarchical clustering plot in PNG format (clustering_RPMK-transformed.png)

6. Enclosures

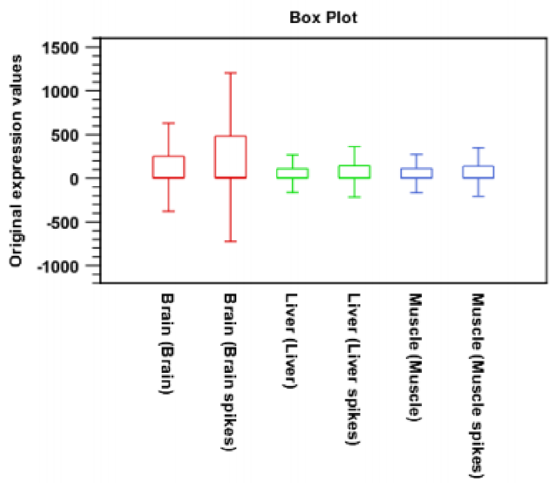
- *Standard:* Summary of the results.
- *Optional*: Quality control of the RNA-Seq experiment based on the overall distribution of expression values in the samples.

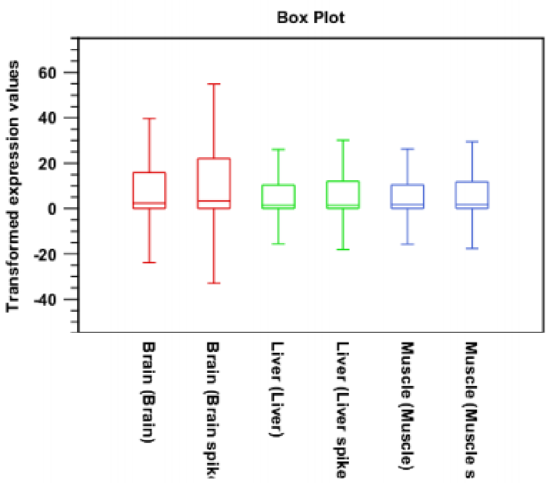
- *Optional:* Explanation of the headers in the RNA-Seq analysis table.

Quality control of the RNA-Seq experiment

Comparison of the overall distribution of RPMK expression values

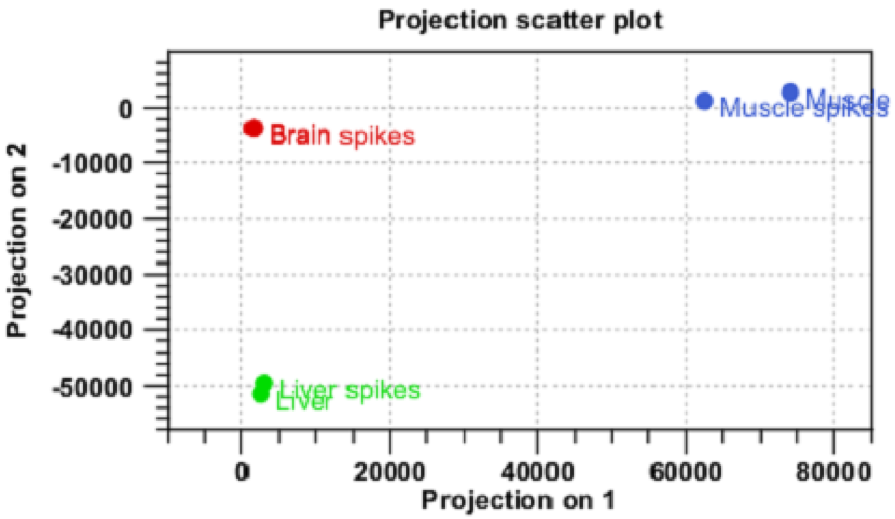
To examine and compare the overall distribution of the expression values in the samples two box plots are generated. These give an overall impression of the locations of the distributions, and to some extend the spread of the distributions. Below two boxplots are shown: the plot on the left is based on the original RPMK values, the plot on the right on the square root transformed RPMK values. The samples are coloured according to the groups.





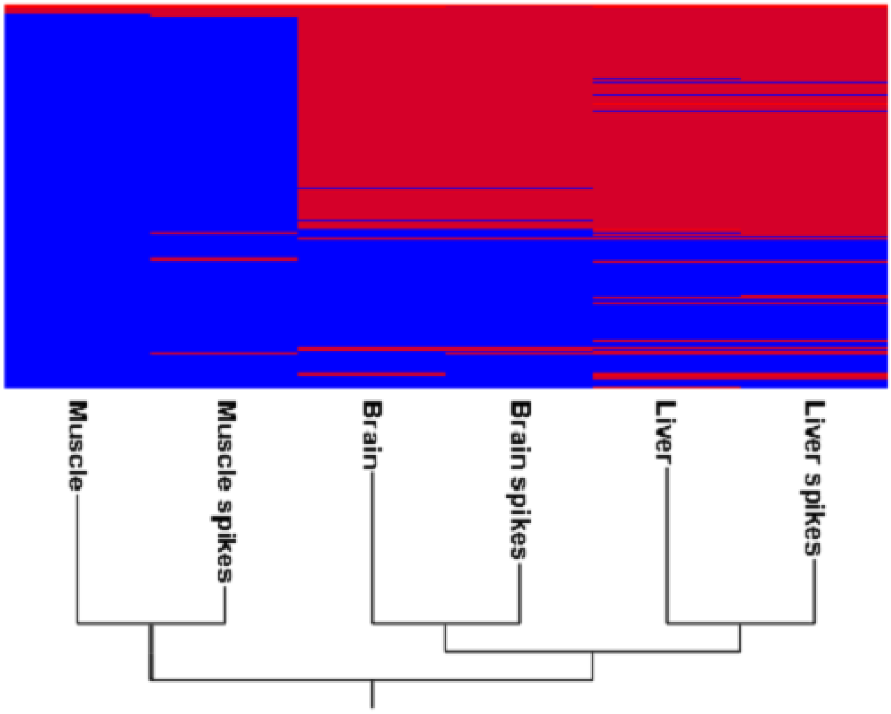
Principle Component Analysis (PCA)

Also the PCA is based on the RPMK values. The plot below shows the projection of the samples onto the two-dimensional space spanned by the first and second principal component. (These are the orthogonal directions in which the data exhibits the largest and second-largest variability).
The dots are colored according to the groups.



Hierarchical clustering

In order to complement the principal component analysis, a hierarchical clustering of the samples to confirm that the samples in the expected groups we expect. Also it gives an overview of the inter- and intra-group variability.



Explanation of the headers in the RNA-Seq analysis table.

Range (original values RPMK).

The 'Range' column contains the difference between the highest and the lowest expression value for the feature over all the samples. If a feature has the value NaN in one or more of the samples the range value is NaN.

IQR (original values RPMK).

The 'IQR' column contains the inter-quantile range of the values for a feature across the samples, that is, the difference between the 75 %-ile value and the 25 %-ile value. For the IQR values, only the numeric values are considered when percentiles are calculated (that is, NaN and +Inf or -Inf values are ignored), and if there are fewer than four samples with numeric values for a feature, the IQR is set to be the difference between the highest and lowest of these.

Difference (original values RPMK).

For a two-group experiment the 'Difference' column contains the difference between the mean of the expression values across the samples assigned to group 2 and the mean of the expression values across the samples assigned to group 1. Thus, if the mean expression level in group 2 is higher than that of group 1 the 'Difference' is positive, and if it is lower the 'Difference' is negative. For experiments with more than two groups the 'Difference' contains the difference between the maximum and minimum of the mean expression values of the groups, multiplied by -1 if the group with the maximum mean expression value occurs after the group with the minimum mean expression value (with the ordering: group 1, group 2, ...).

Fold Change (original values RPMK).

For a two-group experiment the 'Fold Change' tells you how many times bigger the mean expression value in group 2 is relative to that of group 1. If the mean expression value in group 2 is bigger than that in group 1 this value is the mean expression value in group 2 divided by that in group 1. If the mean expression value in group 2 is smaller than that in group 1 the fold change is the mean expression value in group 1 divided by that in group 2 with a negative sign. Thus, if the mean expression levels in group 1 and group 2 are 10 and 50 respectively, the fold change is 5, and if the and if the mean expression levels in group 1 and group 2 are 50 and 10 respectively, the fold change is -5. For experiments with more than two groups, the 'Fold Change' column contains the ratio of the maximum of the mean expression values of the groups to the minimum of the mean expression values of the groups, multiplied by -1 if the group with the maximum mean expression value occurs after the group with the minimum mean expression value (with the ordering: group 1, group 2, ...).

Thus, the sign of the values in the 'Difference' and 'Fold change' columns give the direction of the trend across the groups, going from group 1 to group 2, etc.

The columns under the 'Experiment' header are useful for filtering purposes, e.g. you may wish to ignore features that differ too little in expression levels to be confirmed e.g. by qPCR by filtering on the values in the 'Difference', 'IQR' or 'Fold Change' columns or you may wish to ignore features that do not differ at all by filtering on the 'Range' column.

The RKPM Expression measure

The selected expression measure is the RKPM. It is defined as the Reads per Kilobase of exon model per Million mapped reads (Mortazavi *et. al*, 2008). In other words, it is given by dividing the total number of exon reads (in this case one exon per reference sequence) by the number of mapped reads (in Millions) times the exon length (in this case the length of the reference sequence).

Quality analysis of FASTQ sequence reads

BASECLEAR
FOR 100% DNA RESULTS

project 54114
Client Florida Atlantic Harbor
Date 07-feb-2013

1.

Quality analysis of FASTQ sequence reads

Quality statistics after CASAVA and FastQC analysis

	Number of reads	Sample Yield (in MB)	Average Quality scores (Phred)
25ppt	37,851,786	3,860	38.51
30ppt	14,935,436	1,522	38.49
35ppt	36,093,812	3,680	38.52
bod	37,447,235	3,818	37.75
buc	34,077,063	3,474	37.7
hoc	37,993,699	3,874	37.84

RNA-Seq Analysis Report

project	54114
Client	Florida Atlantic Harbor
Date	07-feb-2013

Summary of the results – RNA-Seq Analysis

Sample 25ppt

Reference sequences

Name	Number of sequences	Longest sequence	Number of genes	Number of transcripts
25-30-35_contig-sequences_Trinity	50,318	9,314	41,946	41,946

Mapping statistics, read mapping*

	Number of paired sequences
Counted fragments	47,310,764
- uniquely	38,522,438
- non-specifically	8,788,326
Uncounted fragments	27,856,855
Total fragments	75,167,619

Detailed mapping statistics*

	Uniquely mapped	Fraction	Non-specifically mapped	Fraction
Exon-exon	0	0	0	0
Exon-intron	0	0	0	0
Total exon	38,522,438	0.81	8,788,326	0.19
Total intron	0	0	0	0
Total gene	38,522,438	0.81	8,788,326	0.19

	Mapped	% of total mapped
Exon-exon	0	0
Exon-intron	0	0
Total exon	47,310,764	100.00
Total intron	0	0
Total gene	47,310,764	100.00

** Default counting scheme ('Fragment counts'): A intact pair is counted as one, broken pairs are ignored*

Summary of the results – RNA-Seq Analysis

Sample 30ppt

Reference sequences

Name	Number of sequences	Longest sequence	Number of genes	Number of transcripts
25-30-35_contig-sequences_Trinity	50,318	9,314	41,946	41,946

Mapping statistics, read mapping*

	Number of paired sequences
Counted fragments	17,924,702
- uniquely	14,620,650
- non-specifically	3,304,052
Uncounted fragments	11,715,399
Total fragments	29,640,101

Detailed mapping statistics*

	Uniquely mapped	Fraction	Non-specifically mapped	Fraction
Exon-exon	0	0	0	0
Exon-intron	0	0	0	0
Total exon	14,620,650	0.82	3,304,052	0.18
Total intron	0	0	0	0
Total gene	14,620,650	0.82	3,304,052	0.18

	Mapped	% of total mapped
Exon-exon	0	0
Exon-intron	0	0
Total exon	17,924,702	100.00
Total intron	0	0
Total gene	17,924,702	100.00

** Default counting scheme ('Fragment counts'): A intact pair is counted as one, broken pairs are ignored*

Summary of the results – RNA-Seq Analysis

Sample 35ppt

Reference sequences

Name	Number of sequences	Longest sequence	Number of genes	Number of transcripts
25-30-35_contig-sequences_Trinity	50,318	9,314	41,946	41,946

Mapping statistics, read mapping*

	Number of paired sequences
Counted fragments	44,962,219
- uniquely	36,919,909
- non-specifically	8,042,310
Uncounted fragments	26,682,945
Total fragments	71,645,164

Detailed mapping statistics*

	Uniquely mapped	Fraction	Non-specifically mapped	Fraction
Exon-exon	0	0	0	0
Exon-intron	0	0	0	0
Total exon	36,919,909	0.82	8,042,310	0.18
Total intron	0	0	0	0
Total gene	36,919,909	0.82	8,042,310	0.18

	Mapped	% of total mapped
Exon-exon	0	0
Exon-intron	0	0
Total exon	44,962,219	100.00
Total intron	0	0
Total gene	44,962,219	100.00

** Default counting scheme ('Fragment counts'): A intact pair is counted as one, broken pairs are ignored*

Summary of the results – RNA-Seq Analysis

Sample buc

Reference sequences

Name	Number of sequences	Longest sequence	Number of genes	Number of transcripts
buc-bod-hoc_contig-sequences_Trinity	31,501	8,331	25,505	25,505

Mapping statistics, read mapping*

	Number of paired sequences
Counted fragments	27,654,234
- uniquely	22,017,889
- non-specifically	5,636,345
Uncounted fragments	39,704,360
Total fragments	67,358,594

Detailed mapping statistics*

	Uniquely mapped	Fraction	Non-specifically mapped	Fraction
Exon-exon	0	0	0	0
Exon-intron	0	0	0	0
Total exon	22,017,889	0.80	5,636,345	0.20
Total intron	0	0	0	0
Total gene	22,017,889	0.80	5,636,345	0.20

	Mapped	% of total mapped
Exon-exon	0	0
Exon-intron	0	0
Total exon	27,654,234	100.00
Total intron	0	0
Total gene	27,654,234	100.00

** Default counting scheme ('Fragment counts'): A intact pair is counted as one, broken pairs are ignored*

Summary of the results – RNA-Seq Analysis

Sample bod

Reference sequences

Name	Number of sequences	Longest sequence	Number of genes	Number of transcripts
buc-bod-hoc_contig-sequences_Trinity	31,501	8,331	25,505	25,505

Mapping statistics, read mapping*

	Number of paired sequences
Counted fragments	32,183,902
- uniquely	25,394,751
- non-specifically	6,789,151
Uncounted fragments	41,962,449
Total fragments	74,146,351

Detailed mapping statistics*

	Uniquely mapped	Fraction	Non-specifically mapped	Fraction
Exon-exon	0	0	0	0
Exon-intron	0	0	0	0
Total exon	25,394,751	0.79	6,789,151	0.21
Total intron	0	0	0	0
Total gene	25,394,751	0.79	6,789,151	0.21

	Mapped	% of total mapped
Exon-exon	0	0
Exon-intron	0	0
Total exon	32,183,902	100.00
Total intron	0	0
Total gene	32,183,902	100.00

** Default counting scheme ('Fragment counts'): A intact pair is counted as one, broken pairs are ignored*

Summary of the results – RNA-Seq Analysis

Sample hoc

Reference sequences

Name	Number of sequences	Longest sequence	Number of genes	Number of transcripts
buc-bod-hoc_contig-sequences_Trinity	31,501	8,331	25,505	25,505

Mapping statistics, read mapping*

	Number of paired sequences
Counted fragments	35,664,711
- uniquely	28,955,086
- non-specifically	6,709,625
Uncounted fragments	39,618,048
Total fragments	75,282,759

Detailed mapping statistics*

	Uniquely mapped	Fraction	Non-specifically mapped	Fraction
Exon-exon	0	0	0	0
Exon-intron	0	0	0	0
Total exon	28,955,086	0.81	6,709,625	0.19
Total intron	0	0	0	0
Total gene	28,955,086	0.81	6,709,625	0.19

	Mapped	% of total mapped
Exon-exon	0	0
Exon-intron	0	0
Total exon	35,664,711	100.00
Total intron	0	0
Total gene	35,664,711	100.00

** Default counting scheme ('Fragment counts'): A intact pair is counted as one, broken pairs are ignored*