Bioinformatics

Understanding the biology of morpholino in zebrafish through integrated gene expression analysis

Kristina K. Gagalova*[1] and Jason Lai[2]

[1]Vrije University, Amsterdam, The Netherlands

[2]Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany

ABSTRACT

Morpholino (MO) is a simple and fast gene knockdown technology used in zebrafish for reverse genetics. New studies have brought into question the reliability of the technology which is highly debated today. The proposed study explores large datasets of gene expression profiles in zebrafish embryos that are treated with MO or with a control MO/untreated. The datasets were integrated together by transforming the raw expression profiles and analyzed with different data analysis approaches. Unsupervised learning was used to explore the gene expression data and to find the main sources of variance: supervised machine learning was applied later for finding differently regulated genes between the two groups. The comparison highlighted the upregulation of p53 and Interleukin-8 pathways. This response may be eventually triggered by DNA damage caused by MO. The study furthermore identified the segmentation as the most deregulated stage by the morpholino treatment. Genes such as *mdm2*, *phlda3* and *abcc5*, respectively coding for p53 negative regulator, p53-regulated factor and ABC drug transporter, showed a significant stage upregulation when compared to controls during the segmentation stage.

1 INTRODUCTION

1.1 Zebrafish: a powerful model organism for genetic studies

The zebrafish (*Danio rerio*) has emerged in the past 20 years as a powerful model organism for the study of development in vertebrates and it is currently used to better understand human diseases (Streisinger et al., 1981). Its transparent embryos permit a detailed microscopic observation during the main developmental stages thus allowing an easy phenotypic characterization. In this scenario, large-scale screenings are easily applicable since several morphological traits are detectable *in vivo* through a non-invasive observation.

The association of gene function to vertebrate development is achieved through identification of mutant phenotypes in zebrafish via two main approaches: forward and reverse genetics (Lawson and Wolfe, 2011). The forward genetics approach shown in Figure 1A identifies gene function through the screening of animal populations where random mutations are induced at low frequencies. Carrier fish animals of a particular phenotype are identified and selected for further breeding. The causative mutations are detected at a later stage through deep sequencing methods. This approach has several limitations. For example the probability of receiving a null mutation in small genes is larger than large genes.

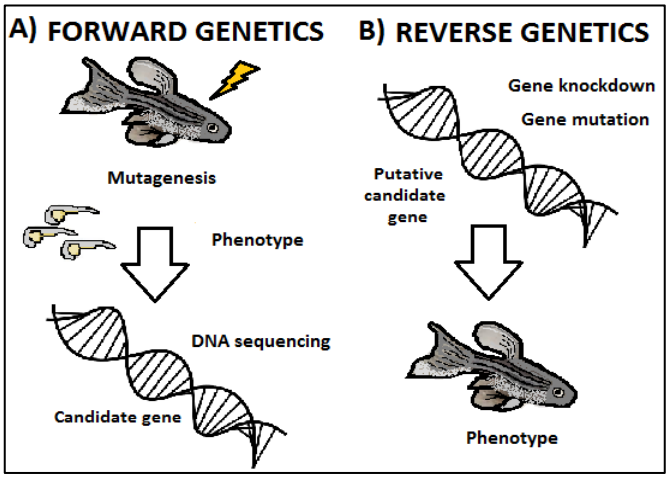

Figure 1 Two main approaches for genetic studies in zebrafish: forward (A) and reverse genetics (B). Forward genetics introduces random mutations and mutants are selectively bred to isolate possible gene mutation after phenotyping. In reverse genetics, candidate genes are targeted for gene inactivation and any resulting phenotype reveals gene function

The recent breakthrough of genomic resources has introduced reliable reverse genetics approaches that directly target the gene or pathway of interest (Figure 1B). In zebrafish, gene inactivation is performed through gene knockdown which transiently inhibits translation or splicing of a target mRNA (Nasevicious and Ekker, 2000; Draper et al., 2002). Another approach may directly mutate those genes where the forward genetics has failed to generate a mutation. This method complements the reverse genetics and aims to enrich the spectrum of available mutations in gene function studies. The reverse genetics approaches offer a direct interpretation of the gene function from the resulting mutant phenotype.

1.2 Making sense with antisense: the Morpholino oligonucleotide

The use of antisense oligonucleotides in reverse genetics screens was adopted with considerable agreement because of its easy design and application. The oligonucleotides are designed to specifically bind to a targeted mRNA sequence and to interfere with the transcript maturation or translation. Several generations of antisense oligonucleotide were created by progressively improving their stability and efficacy. A successful antisense oligonucleotide used in developmental biology is the morpholino antisense (MO) that was initially introduced by J. Summerton and D. Weller (Summerton and Weller, 1997) for therapeutic applications.

Morpholinos are short synthetic nucleic acid analogs, composed of 25 bases chain. Each subunit is comprised of a nucleic acid base, a morpholine ring and a non-ionic phosphorodiamidate intersubunit linkage (Figure 2). Unlike the DNA phosphate backbone, the morpholino antisense uses phosphorodiamidate for the backbone to ensure a neutral charge of the molecule thus avoiding ionization that may occur at physiological pH. The molecule is resistant to degradation by biological agents such as RNAase or DNAase and it is demonstrated to have a relatively low toxicity when used for clinical treatments *in vivo* (Pandey et al., 2014; Takei et al., 2005). Heteroduplex denaturation assays (Janson and During, 2006) show that morpholinos have high affinity for RNA which allows the synthetic molecule to efficiently invade the RNA secondary structures.

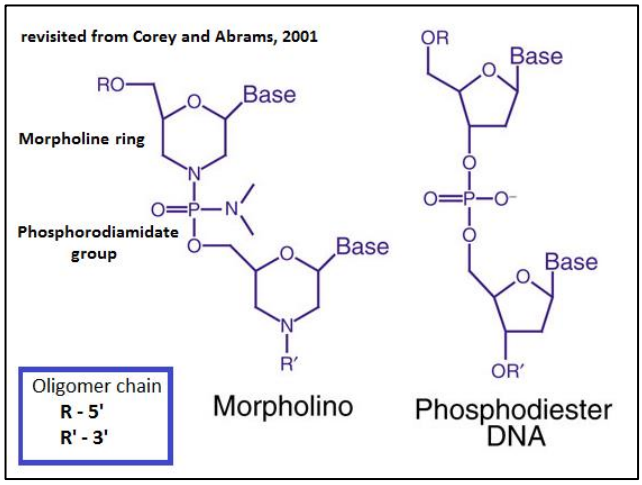

Figure 2 Morpholino oligo compared to DNA. The phosphorodiamidate group in morpholino is not ionizable in biological environment. The morpholinos oligo chemical composition allows an efficient binding to the target RNA and resists degradation from biological agents.

Two morpholino gene inactivation strategies are employed in zebrafish (Figure 3). The first mechanism occurs in the cytoplasm and prevents mRNA translation by inhibiting the progression of the ribosomal initiation complex by steric hindrance: in this case the translation-blocking morpholino targets the 5´ untranslated region (5´- UTR) of the mRNA (Figure 3A). The second mechanism occurs in the nucleus and prevents transcript splicing by blocking small nuclear ribonucleoproteins complexes from binding to their targets at the exon-intron boundaries (Figure 3C). The unspliced mRNA could be degraded from one of its ends by activating molecular pathways in the cytoplasm that eliminates mRNA with particular structural defects, called "Nonsense mediated decay" (Ward et al., 2014) or "Non-stop decay", based on the entity of the splicing deregulation (Baker and Parker, 2004). The translation-blocking mechanism on the other hand is known to not cause mRNA degradation through the previously described molecular mechanisms: the target mRNA is trapped by the morpholino molecule in the cell cytoplasm. Moreover morpholino can efficiently block also micro-RNA activity too (Figure 3B).

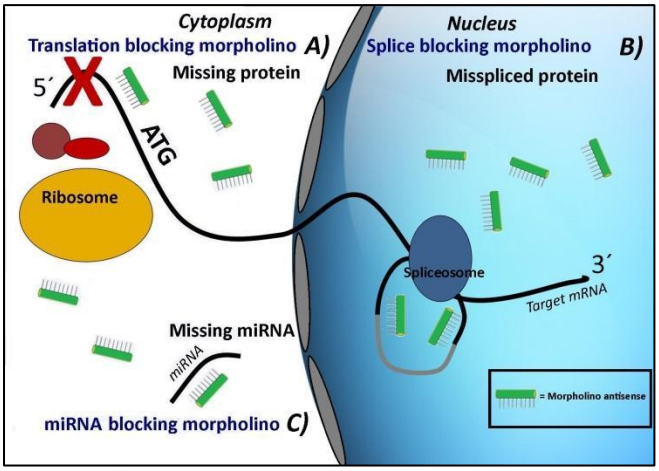

Figure 3. Different inhibition mechanisms of morpholinos in zebrafish. The translation-blocking morpholino (A) targets the 5´ UTR region of the mRNA and inhibits protein translation. The splice-blocking morpholino (B) binds to the splicing junctions and interferes with transcript maturation. Morpholinos can also be designed to target miRNA molecules (C).

The morpholino is injected in Zebrafish embryos at 1-8 cells stage and the molecules homogeneously spread through the cytoplasmic bridges connecting the cells (Bill et al., 2009). The morpholino dosage varies from gene to gene and it is usually determined by increasing the morpholino amount before inducing undesired and deleterious effects in the embryos. Several molecular techniques are used to assay its efficacy such as detecting the absence of the targeted protein or amplifying the unspliced gene variant.

An insidious problem in the use of morpholinos is the possible presence of collateral and artefact effects. The surplus of the morpholinos in the cell environment not only functions to knockdown a target transcript but may also give rise to unexpected complementarity to other poorly annotated genes or to nonspecific effect. The latter, usually referred as "off-target effect", is due to the interaction between the oligo and the extracellular, cell-surface or intracellular proteins structures (Summerton, 2007). Until now the only molecular mechanism described as nonspecific effect is the activation of the p53 cell death pathway (Robu et al., 2007). About 15-20% of the morphants shows a particular set of off-target effects, visible during the segmentation phase of development. However, there has been little empirical research for determining the side effects mechanisms of the morpholinos *in vivo* and little is still known about the mechanisms which increase the activation of the p53 pathway.

The most broadly used control strategy which takes into account off-target effects is the use of control morpholinos which do not interfere with the expression of any endogenous gene. For this purpose there are several possible strategies applied in the morpholino studies. One strategy is the use of a "Standard Control" which is designed to target a human beta-globin intron variant than

causes beta-thalassemia (GeneTools, 2015). This experimental design does not necessarily control for the activation of the p53 protein so it is usually recommended to use a similar to the morpholino sequence which however has several mismatches that guarantee a suitable control (mismatch control) (Eisen and Smith, 2008). Other control strategies involve a scrambled morpholino sequence, the injection with buffer solution or, in many cases, the mere uninjected zebrafish embryos. The use of control for morpholino knockdown is still poorly regulated and the morpholino users rely on different approaches for their experiments.

1.3 Some gene mutations does not recapitulate the morpholino phenotype

New genome editing tools have recently become available (Cermak et al., 2011; Cong et al., 2013; Mali et al., 2013). Site specific nuclease has been extensively used to edit the genomic DNA in zebrafish. Several developmental research groups have put significant effort in the improvement of those techniques, applied in parallel with the oligo antisense morpholino in reverse genetics. These new technologies have made it possible to conduct a systematic comparison of the phenotypes generated between knockdown and mutation techniques for a given gene.

A recent report by Kok and colleagues (Kok et al., 2015) have analyzed a set of mutant lines for genes involved in the embryonic development and have compared the observations to their respective morphants. A subset of 20 genes emerged as not sharing the same phenotype when compared between mutants and morphants. The authors of the paper explained the observed discrepancy as due to the off-target effects of morpholino but other recent research streams consider the occurrence of specific cellular mechanisms that may cause the missing correspondence. A recent study by Rossi, Kontarakis and colleagues (Rossi, Kontarakis et al., in press) compared mutants and morphants of Epidermal Growth Factor Like Receptor – 7 (*egfl7*). The study aimed to identify the molecular mechanisms underlying the difference in phenotypes. The proteome data analysis highlighted the upregulation of Extracellular Matrix proteins (ECM) which may be involved in compensation mechanisms occurring in mutant animals. Both the studies highlighted however the possibility of a higher than expected off-target effects in morphant animals. Even if several morpholino oligos are successfully exploited for gene silencing, further information is required for the conscious application of the morpholino antisense in zebrafish reverse screenings.

1.4 Machine learning: making sense of integrated gene expression profiles

The Gene Expression Omnibus (GEO) (Edgar et al., 2002), maintained by the National Center for Biotechnology Information (NCBI), is a repository that collects datasets from different high-throughput technologies. Several gene expression profiles from publications involving morpholino gene knockdown have been shared on GEO, mainly generated with microarray technology. The data sets compared the gene expression of animals injected with morpholinos to the corresponding controls. Even if those data sets were produced for addressing different questions, the combined analysis of these data is capable of highlighting common patterns (Rhodes et al., 2002) and to answer complex biological questions such as the possible off-target effects of morpholinos.

A successful method for analyzing an inter-study microarray is the direct integration of gene expression profiles: in this approach the microarrays are combined by transforming the raw expression values through normalization techniques. This kind of data analysis does not allow the direct comparison of the genes between treatment and control by simply averaging the probesets. Typically gene expression data suffer from limited number of samples and high dimensional data properties. This problems are overcome by statistics and machine learning methods which are already successfully applied in cancer biomarkers detection (Jagga and Gupta, 2014; Lee et all, 2011). The main objective of those studies is to find a small set of biomarkers with strong discriminant power between the groups that can be also used for cancer subtypes prediction. The task of selecting informative biomarkers from microarray datasets has already been studied with several feature selection methods and classification models (Hsu and Lu, 2008; Pirooznia et al., 2008).

The current study proposes an integrated microarrays data analysis of public available morpholino datasets. The main goal of the study is to identify possible nonspecific molecular interactions of the morpholino antisense *in vivo* and to determine the molecular pathways involved in the off-target morpholino effects. As first approach the data were submitted to unsupervised learning which gave a general overview of the data sets. After the identification of the main sources of variability, the data sets were explored through supervised machine learning methods for detecting common patterns of gene expression in the morpholino treated samples.

2 METHODS

2.1 Data sets selection and normalization

2.1.1 Microarray data selection and curation

The gene expression profiles characterizing morpholino injection in zebrafish are retrieved from GEO data repository (GEO, 2015) which is browsed on the 19th March 2015. The key-words used for the search are “zebrafish morpholino” and “zebrafish knockdown”: the filters are restricted to “Danio rerio” as model organism and “GEO series” as entry type. The hits resulted from the search which effectively refer to the entered key-words are 36 that correspond to the number of studies with deposited microarray results. Diverse technologies cannot be combined due to different hybridization properties and intensities normalization strategies thus only the topmost frequent microarray technologies are selected for the study, namely Affymetrix and Agilent. The id of the two technologies platforms are shown in Table S2 with the corresponding number of samples. These technologies are used for creating two separated microarray datasets. Different Agilent platforms are integrated when possible for increasing the number of samples in the study. Since each Agilent platform is designed for a specific set of genes, only the studies with the highest possible overlap are combined together. For this purpose, the complete list of gene ids from each platform is intersected with the other Agilent platform gene ids. An example of the output from the gene overlap study is shown in the Venn diagram in Figure S2.1.1A where the highest number of overlapping genes is obtained with microarray platforms 14664 and 6457. Customized Agilent microarrays share lower number of genes with the other platforms (Figure S2.1B). The microarray samples used for data analysis are those from Affymetrix GPL1319 platform, comprehensive of 104 instances and 11516 probes mapping unique genes, and Agilent microarray samples from

GPL1464 and GPL6457, containing 89 instances and 13856 probes mapping unique genes. The Tables S2.4 and S2.5 contain detailed information of the microarray samples considered in the current study.

The information from the selected studies is parsed through R package GEOquery (Davis and Meltzer, 2007) and fetch_GSEInfo.R script (***1A***). The datasets are manually curated and detailed information about the experimental design is extracted from the annexed scientific literature. Microarray data sets without reference to scientific reviews are discarded from the study. Every microarray profile is characterized by the following attributes: GPL, GSE, GSM, Title, Hours post fertilization, Source, MO dose, Treatment, MO sequence, Control (if available) sequence, MO target, MO type. The hours post fertilization attributes is converted in categorical variable by assigning a developmental stage to the corresponding time point. Furthermore, the ng of injected morpholino is converted to picoM when possible for summarizing the dose and the MO sequence in one unique quantity. The missing morpholino dose values are retrieved from analogous gene silencing studies when possible. More information about the data set curation is available in the Table S2.2 and Figure S2.3 of the supplementary material.

2.1.2 Intensities normalization and processing

The row intensities are combined together and normalized according to standard microarrays normalization protocols. This approach aims to settle a meaningful comparison of the expression levels.

Affymetrix: the samples are normalized with Single Channel Array Normalization (SCAN) (Piccolo et al., 2012) which is defined a standardization technique applied individually to every microarray sample. SCAN utilizes a modification of the Model-based Analysis of tilling-arrays (Johnson et al., 2006) which is based on the individual probe sequence signal normalization. The new SCAN method calculates the effect of two Gaussian distributions compounded by signal from the background noise and the biological effective noise. The output intensities are corrected by the background signal and standardized on the estimated variance of probes with similar expected background. In this normalization strategy only the perfect match probes (PM) are considered while the mismatch (MM) probes are ignored. The Affymetrix data are read with *ReadAffy* function in *affy* R package. The normalization is performed with SCAN function used through NormalizeAffy.R script (***2A***).

Agilent: The normalization is executed on both the Single and Double channel Agilent technologies. The intensities are treated as independent samples in the data integration. The background correction is performed with "minimum" method correction due to the high number of negative intensities. The method transforms the negative values after the background normalization to half minimum of the positively corrected intensities. The output values are normalized with Variance Stabilization Normalization (VSN) (Huber et al., 2002). Since the variance is observed to increase considerably with the intensity of the signal, the normalization applies an intensity variance correction that is maintained constant across the whole intensity range. Also in this case the normalization is applied singularly for every sample. The R package used for normalization and for reading the Agilent files is *limma* (Smyth et al., 2011). The normalization is performed through the *normalizeVSN* and *backgroundCorrect* functions in NormalizeAgilent.R (***2B***).

The output data for both the microarrays technologies are corrected for "batch" effects which is the high variability introduced with noise not imputable to biological conditions. The approach here used is based on the locate and scale adjustments (L/S) where the data are fist standardized for common mean/variance and then adjusted according to Empirical Bayes (EB) parameters (Johnson et al., 2006). The data are adjusted for the estimated EB batch effects calculated on the studies batch covariate. The R package which implements this method is *sva* used through ComBat function (Combatting Batch effects when combining Batches of Gene Expression Microarrays data) in the NormalizeAffy.R and NormalizeAgilent.R scripts. The boxplots in the supplementary information show the intensities distribution for the row (S2.5A, S2.6A), the normalized (S2.5B, S2.6B-C) and the batch corrected intensities (S2.5C, S2.6D) for Affymetrix and Agilent. As possible to see the intensities are first corrected for the technical microarray variance, such as the nucleotide content or background, and finally smoothed for batch effect.

Control probes are removed from the data sets and probes hybridizing to the same gene product are averaged. The data are filtered for removing the low variance genes: the method used for the pre-filtering statistics is the inter-quantile range and values lower than 0.6 variance cut-off are removed. The function used for the purpose is *varFilter* from the *geneFilter* R package. The output score are additionally filtered for the genes showing high positive correlation in the same microarray sample. The Affymetrix data set is filtered for values with Pearson correlation higher than 0.95 and the Agilent for values higher than 0.9. Correlation filtering is carried out through *findCorrelation* function from *caret* R package. The final data are submitted to centering and scaling. The data filtering and scaling is performed with NormalizeData.R script (***2C***). The data processing and normalization described before is shown in the workflow in Figure 4.

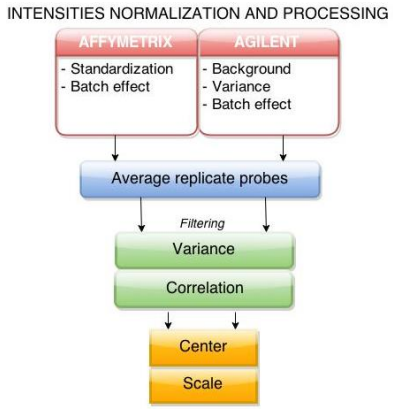

Figure 4 Microarrays intensities processing. The normalization strategies are applied based on the technology characteristics. The data processing filters redundant and poorly informative features. The last processing step centers and scales the output data.

2.2 Data sets exploration and factors interaction

2.2.1 Unsupervised machine learning

The data sets are explored through standard PCA and hierarchical clustering for inferring the existence of possible samples grouping. The PCA is performed through PCA.R (***3A***) and clustering trough HierarClustering.R (***3B***).

2.2.2 Linear model and ANOVA

The data sets are investigated for possible interaction between the morpholino treatment and the developmental stage trough linear ANOVA decomposition of Multivariate Design. This approach aims to determine the different stage response effect that may arise after morpholino treatment. The gene interaction is visualized through biplot (Gabriel, 1971) and also shown as PC1 loadings plot per stage. The ANOVA decomposition which is based on ANOVA-Simultaneous Component Analysis (ASCA) (Smilde et al., 2005) considers the factor levels of the two variables (developmental stage and treatment) together with the replicates for every interaction level. In the current analysis the model factors are estimated by the *maximum likelihood* approach, using the lmFit function provided by *limma* package. The data visualization is performed with PCA decomposition. The data analysis is executed with linRegressionModel.R script using lmdme R package (***3C***).

2.3 Supervised machine learning

2.3.1 Feature selection models

Different feature selection approaches are compared in order to define a suitable model which could be used to distinguish differently expressed genes in the MO when compared to the control. All the samples were included despites the tissue origin, the developmental stage or the morpholino dosage injected.

The strategies here used for detecting differently expressed genes a group of particular filtering and embedded feature selection methods which are shown to perform well on high dimensional datasets such as microarrays (Saeys et al., 2007).

Embedded feature selection: in this approach the features interact with the learning algorithm and features search is guided by the learning process itself. The choice of the learning algorithms is made for covering a broad range of decision boundaries in the microarrays datasets. The embedded machine learning algorithms recognize linear boundaries trough linear SVM, Gaussian process, General Linear model with elastic net, quadratic decision boundaries through Random Forest, J48 and Naïve Bayes and more complex boundaries through radial SVM, polynomial SVM and partition around medoids. The prediction performance is estimated through Leave One Out Cross Validation (LOOCV) and accuracy score. The learning parameters are optimized singularly for every learning algorithm. The algorithms are used trough *train* function in *caret* R package and the following methods in RunEnbedded.R script (***4A***): "*svmLinear*", *"gaussprLinear", "glmnet", "rf", "J48", "nb","svmRadial","svmPoly", "pam"*.

Filtering feature selection: this approach measures the importance of the features subset based on the intrinsic properties of the dataset. The filtering feature selectors are independent of the proposed model and are usually divided in univariate and multivariate filter methods. While the univariate filtering methods scores individually the importance for each feature, the multivariate methods delve in more complex feature interactions such as the feature correlation. The filter feature selectors used in the study are the t-test for the univariate and the ReliefF (Robnik-Šikonja M, Kononenko, 1997) and the minimum redundancy maximum relevance (mRMR) (Ding and Peng, 2005) for the multivariate. The t-test feature selector is implemented through 4-folds stratified sampling without replacement. The t-test is performed on ¾ of the original instances and the genes significant at p-value 0.05 are assigned to a score. Higher is the gene significance (lower p-value) and higher score is assigned to the gene. After one run of resampling where all the samples are used for t-test, the 4 list of scored genes are averaged. The t-test resampling is iterated *n* times and gene scores is summed together at the end of the loop. Additional information on the t-test feature selection could be found in the Figure S2.8. The feature selectors are tested trough different learning algorithms for validating their performance. The K-nearest neighbor (KNN) algorithm is used for reference for the filter methods. The t-test feature selector was performed trough Ttest.R script (***4B***) and the ReliefF and mRMR through Rweka and mRMRe R packages, used in MultivariateFeature.R script (***4C***). The feature relevance is evaluated by adding an increasing number of features and validating the model performance through LOOCV. The highest accuracy pick reached by the model is used for comparison with the embedded methods. The ranked features are evaluated trough *knn.cv* function from *class* R package and knnValidation.R script (***4D***).

The actual labels model is compared to a set of 5 models which predictor labels are permutated (scrambled training set). The scrambled training sets are evaluated trough LOOCV and compared to the actual labels trained model for control.

2.3.2 Optimal feature number selection

The best performing model is selected based on the highest accuracy calculated trough LOOCV and on the distance to the average accuracy of scrambled training sets. The number of features to be selected is decided on the accuracy contribution cumulative features. The number of optimal features is selected differently in the embedded and filtering feature selectors.

Embedded feature selection: The features are evaluated trough recursive feature elimination and backward selection. The model is trained using all features and its overall performance is estimated trough 4 folds cross-validation repeated 1000 times. The algorithm uses different subsets of features (sizes) that are evaluated and selected from the all-features data set based on the importance covered in the model. The model is trained with the selected features and the performance is evaluated with the schema mentioned before. Since this approach is prone to overfitting an additional cross-validation loop is introduced while incorporating resampling. The features subset evaluation is performed for 4 fold partition of the instances and averaged over 100 iterations. A detailed description of the algorithm loop is shown in the Figure S2.7. Coarse feature number accuracy estimation is calculated for broad feature sizes intervals (1, 500, 1000, 1500…, 3768) which gave an overall distribution of the accuracy for feature number. Since the highest accuracy score is achieved with the all-features data set a range of optimal features is selected with an accuracy tolerance interval of 5%. The feature number accuracy is then calculated for narrower interval (120, 121, 122, 123,…150) and the optimal number of features is finally selected for being closer to the tolerance threshold. The recursive feature elimination is used with rfe function in caret R package and FeatureNumberEmbedded.R script (***5A***).

Filtering feature selection: The number of features is increased additively (1, 2, 3,…. 3768) and the LOOCV prediction accuracy is calculated for every feature step in the KNN model. The accuracy estimation is performed for both the scrambled and actual labels training sets. The accuracy distribution across the different number of features is smoothed with LOcal regrESSion (LOESS). The LOESS fit is optimized through improved Akaike Information Criterion and FitAIC_LOESS.R script (***5B***). The number features is selected in correspondence of the highest difference between the actual labels and the scrambled training sets estimated trough knnOptimalFeatures.R script (***5C***). The Figure 5 shows the general approach used for the feature selection in the filter and embedded methods.

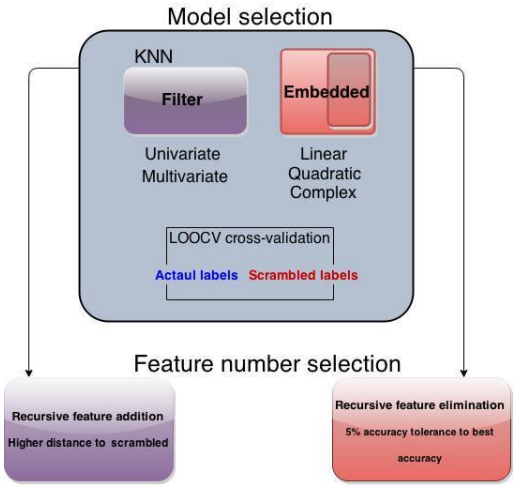

Figure 5 Feature selection strategy. Several embedded and filter feature selection methods are evaluated trough LOOCV and the best performing method is selected. The optimal number of features is decided through recursive feature addition and elimination for respectively the embedded and filter methods. A set of scrambled label training models are used for avoiding overfitting and for feature number selection for the filter feature selection methods

2.4 Pathway data analysis and gene expression enrichment

The genes selected with the optimal feature selection model are analyzed with Gene Set Enrichment Analysis (GSEA) for GO terms molecular function and biological process (The gene ontology consortium, 2000). The GSEA is performed with GOrila - Gene Ontology enRIchment anaLisis and visuaLizAtion online tool (Eden et al., 2007). The analysis is carried out on the ranked list of features according to the machine learning model. The genes are compared to both Human and zebrafish Gene Onthology data base for relaying on a more detailed gene annotation. The GSEA is run at standard parameters.

An additional Ingenuity Core pathway data analysis is executed in IPA-QUIAGEN (Qiagen, 2015) for observing the gene deregulation on pathway level. This study considers together with the significance score from the classical GSEA also the direction of gene regulation (upregulated or downregulated).

3 RESULTS

3.1 Data exploration

The two datasets were initially explored with standard Principal Component Analysis (PCA) and Hierarchical clustering. The PCA of the Affymetrix dataset in Figure 6 highlighted the main sources of biological variance, namely developmental stage and tissue type.

The gene expression is strongly influenced by the developmental stage which represented 29% of the PCA variance observed in first principal component (PC1). The figure distinguished between the early and the late developmental stages which are separable in the PCA plot. The secondary source of variation in the data set was the tissue type: gene expression from whole embryo and single tissue samples are separated by PC2. Gene expression characterizing the heart samples was closer to the endothelial cell samples; the gene expression from kidney tissue was well grouped too.

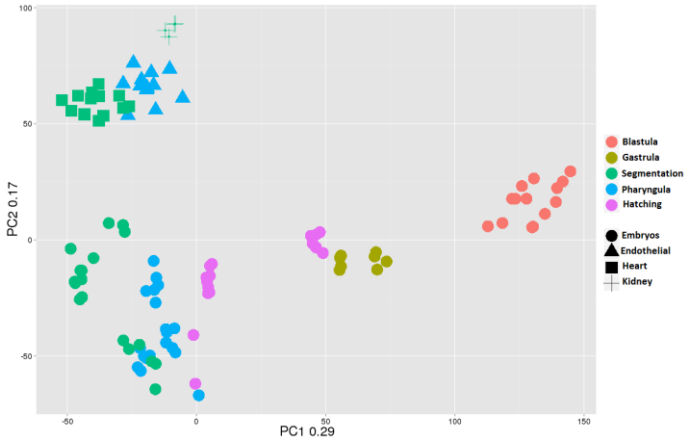

Figure 6 PCA of Affymetrix gene expression data set. The samples are colored based on the developmental stage (Blastula: 2.25 -5.25 hpf, Gastrula: 5.25-10.33 hpf, Segmentation: 10.33-24 hpf, Pharyngula: 24-48 hpf, Hatching: 48-72 hpf). The symbols show the tissue origin of the zebrafish samples. Hpf- hours post fertilization

The PCA was additionally labeled based on the different type of treatments: uninjected (WT), injected with control morpholino (CO) and injected with target morpholino (MO) (Figure S3.1). However no clear grouping was observed for the three classes after inspecting the PCA output for the first five main principal components (Figure S3.1A, B, C, D). These results were confirmed also with hierarchical clustering where the tissue type was the main clustering factor among the samples (Figure S3.2C). The gene expression pattern of morpholino injected samples was not appreciably distinct from control samples. The dataset was additionally labeled for the different dose of injected morpholino which was however also missing of clear group separation (data not shown).

The Agilent dataset was also explored through PCA shown in the Figures S3.3A, B, C and D in the supplementary material. Unlike the previous dataset, the data exploration showed poor clustering. The main source of variation was expected to be given mainly by the developmental stage since the RNA samples in the data set were only extracted from whole embryos and not from specific tissues. High number of samples was concentrated in the central region of the plot where the variance was poorly explained by the PCA data transformation. Thus only the Affymetrix data set was further analyzed.

3.2 Feature selection: differentially expressed genes in morphants

3.2.1 Feature selection model

To gain a deeper insight in the morpholino treatment, a supervised machine learning approach was adopted to define specific candidate genes that are differentially regulated in the morphants. Feature selection was first applied to define a subset of potential biomarkers that characterize the morpholino treatment. The feature selection was improved and compared across different methods which aimed

to build a solid model for identifying valuable biomarkers. Table 1 shows the feature selection methods and the correcponding LOOCV accuracy for the Affymetrix dataset. Each LOOCV was compared against a set of scrambled models as a random baseline. Similar LOOCVs between the actual and scrambled models indicate that the selected features were not more informative than a random guess.

Table 1A Feature selection – embedded methods

Feature selector	LOOCV	LOOCV scrambled
Glmnet	0.913	0.488 ± 0.045
Gaussian Linear process	0.933	0.565 ± 0.076
SVM Linear kernel	0.933	0.565 ± 0.076
Random Forest	0.836	0.542 ± 0.087
J48	0.798	0.509 ± 0.041
Naïve Bayes	0.57	0.561 ± 0.065
SVM Radial kernel	0.663	0.563 ± 0.1053
SVM Polynomial kernel	0.875	0.613 ± 0.063
PAM	0.635	0.512 ± 0.057
KNN	0.644	0.534 ± 0.075

Table 1B Feature selection – filter methods and KNN

Feature selector	Max LOOCV	Max LOOCV scrambled	Max distance to scramble
t-test	0.732 (86 feat.)	0.644 (501 feat.)	0.722 (1 feat.)
Relief	0.781 (135 feat.)	0.602 (3550 feat.)	0.781 (130 feat.)
mRMR	0.588 (269 feat.)	0.638 (377 feat.)	0.534 (1 feat.)

The maximal LOOCV accuracy score was obtained with learning models recognizing linear boundaries. The best performing models were the linear kernel SVM with the regularization parameter C equal to 1 and Gaussian Linear process. The Gaussian Linear model was in this case preferred because of its automatic tuning and absence of model parameters. The Agilent data set showed higher prediction accuracy with the filter feature selectors (Table S3.4B). The highest accuracy was reached with t-test and K-nearest neighbor (KNN) model. In order to prove that the accuracy scores are not uniquely given by the model used for testing the filters, the KNN was also evaluated as embedded feature selector. The accuracy score was higher after using filters feature selectors thus showing the improvement given to the model.

The mRMR filter feature selector had low prediction accuracy and its performance was similar to the scrambled label models. The algorithm computes the mutual information and the correlation score given a starting set of relevant features (seeds). mRMR´s performance was tested with one or ten seeds: the filter feature selector was greatly improved by multiple starting features which was however computationally expensive (data not shown). The algorithm here tested was considered not suitable for large dimensional data sets such as microarrays. The parameters used in the embedded methods are shown in Table S3.5. The parameters were optimized for reaching the highest possible accuracy in the prediction.

The number of features that were selected from the best performing method in the Affymetrix dataset was defined with recursive feature elimination and 5% tolerance of the maximum accuracy score. The process of selecting the number of features is shown in Figures S3.6A and S3.6B. A broader interval of features was chosen for defining the rough estimate of features around the tolerance interval. The exact number of features was determined after the estimation of accuracy in the narrower interval of features.

3.2.1 The biology behind the selected biomarkers: Gene Enrichment Analysis and Ingenuity Pathway Analysis

The genes selected through feature selection were additionally studied for their biological properties. The full list of selected genes is included in Table S3.7 together with the type of regulation and corresponding score assigned by the Gaussian Linear model. The ranked genes from the model were annotated with 'Molecular Function' and 'Biological Processes' (Figure 7A and 7B) Gene Ontology (GO) (The Gene Ontology Consortium, 2000) terms and ranked on the score assigned in the model.

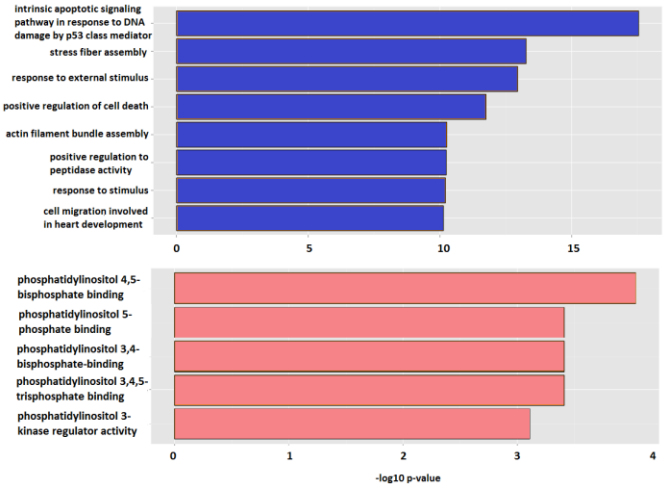

Figure 7. (A) Significant Molecular function and Biological process **(B)** terms in GSE analysis

The gene ontology terms obtained from GSEA highlight two main properties of the morpholino treatment. The morpholino antisense bound to the mRNA triggers several cellular processes such as response to external stimulus and the assembly of stress fibers. The other biological effect induced in morphants is the activation of the apoptotic pathway through the intrinsic signaling pathway and the DNA damage. The molecular function highlights the activation of signaling proteins in the plasma membrane. More detailed information about GSEA analysis is shown in Tables S3.8A and S3.8B.

Figure 9 shows the pathway analysis from IPA Ingenuity software.

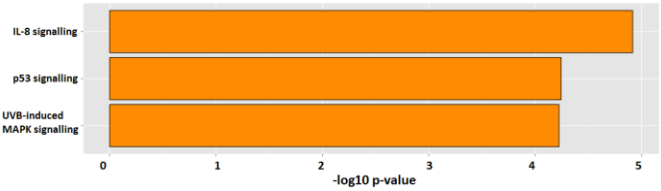

Figure 9. IPA pathway analysis – significantly upregulated pathways in morphant animals.

The IL-8 and p53 signaling pathways are activated in morphants; the former indicates an activation of the immune system, while the latter of apoptotic or DNA damage pathways.

3.3 Developmental stage and treatment: interaction effects

The gene expression profiles were analyzed with two-way analysis of variance (ANOVA) for multivariate levels of the microarray samples. 15 genes resulted significant when tested with level of α equal to 0.001 for the interaction effect thus showing a significant stage differential expression in the two treatments.

The 15 candidate genes were used in a Biplot (Figure S3.9A) to highlight the contribution of every single gene to the possible interaction terms (stage and treatment) where the interaction is represented as a vector in the space of the two principal components. The figure showed that the significant genes are mainly located in proximity to segmentation and hatching stages.

In order to gain a more detailed view of the stage effect, the loadings of the PC1 was displayed for factor level. Figure S3.9B showed the interaction effect of two consecutive levels of stage factors and the strong interaction effect that was mainly observed during the segmentation stage.

The median gene expression levels of the significant genes were shown by developmental stage for the morpholino and control treatments in Figure 10.

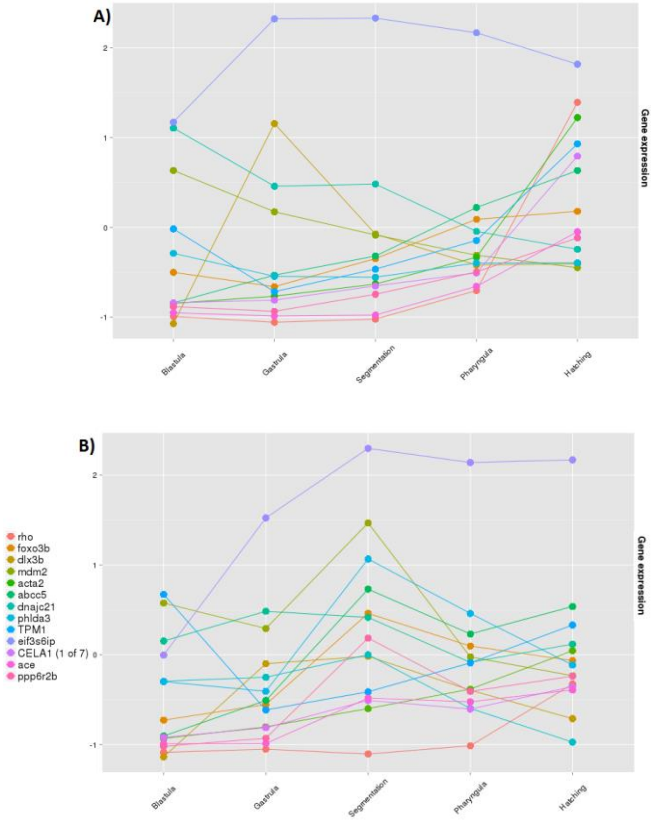

Figure 10. Genes with significant interaction between treatment and developmental stage. The level of gene expression is shown for the genes significant at p-value 0.001. The figure A shows the gene expression for the control treatment and the figure B for the morpholino treatment. The genes are shown by gene symbol.

The complete list of significant genes found can be found in the Table S3.10. The genes shown in the figure have a different expression pattern in the two treatments: the gene expression in the morphants generally shows a higher gene expression effect in the segmentation stage. Particularly interesting is the gene expression for a subset of genes that show higher gene expression during the hatching stage in the control treatment, composed of standard control, buffer injected and uninjected samples. The same genes have an earlier peak of gene expression level during the segmentation stage. The effect described before is particularly evident for mdm2, *abcc5*, *acta2* and *phlda3* genes.

As previously reported, *p53* is commonly upregulated in morphants (Robu et al., 2007). Thus the gene expression levels of *p53* were plotted for both morphants and control against various developmental stages. Figure 11 shows the gene expression level by stage.

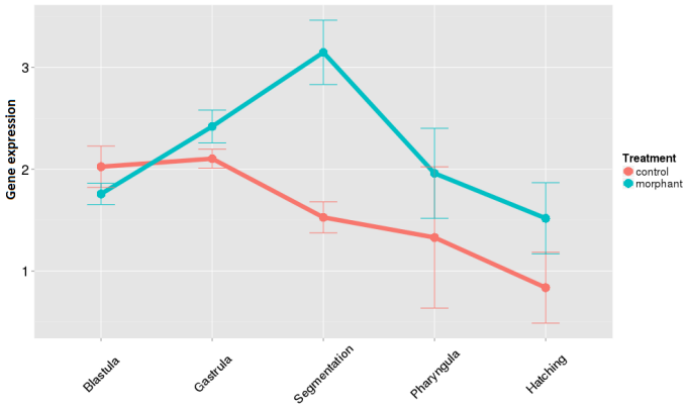

Figure 11. *p53* gene expression level per developmental stage

p53 was shown to be significantly upregulated in the segmentation stage as well. The upregulation was significant at p-value 0.05 when considering the interaction effect in linear model of designed multivariate experiments.

4 DISCUSSION AND CONCLUSIONS

4.1 Datasets processing and exploration

The current data analysis aimed to detect the possible off-target effects of morpholino knockdown to study developmental genetics. The study started with the data selection from GEO gene expression repository. The data were selected taking into account the technology compatibility and the platform features. Two integrated gene expression data sets were produced from the most common and frequently used technologies, Affymetrix and Agilent. The Affymetrix is a high density short oligo array (25 base pairs) while the Agilent uses long oligo nucleotides for the hybridization (60 base pairs). The Affymetrix microarray technology utilizes a single intensity fluorescence dye; the Agilent instead uses two different dyes that emit fluorescence in the green and in the red spectral intensities. The described technical features for the two technologies have different physical and chemical properties which could be recognized in different platform variability and sensibility (Carole et al., 2004). The probes gene complementarity is moreover different between microarray platforms even if hybridizing to the same gene because of their length and design. Each normalization strategy was decided on the main probe biases. The processed data were normalized for batch effect and intrinsic intensity variance.

The normalization was inspected with PCA and hierarchical clustering for identifying the main underlying variance in the pooled microarray data. The Affymetrix dataset highlighted a clear data pattern where the tissue source and the developmental stage were the two main factors that explain the PCA grouping pattern. The vertebrate's embryogenesis is one of the most complicated developmental programs and several critical events such as cleavage, blastulation, gastrulation and somitogenesis are regulated by striking genome-wide transcription reprogramming (Yang et al., 2013). The developmental gene expression is therefore expected to be dramatically different between microarrays together with individual tissue-specific gene expression (Liang et al., 2006) which together contribute to the main biological variance in the dataset. In contrast, the PCA from Agilent had no clear factors to explain the majority of variance in the dataset. The PCA plot was inspected in the first five principal components for finding further variability. The lack of clustering by developmental stage and tissue type in the Agilent dataset is mostly likely to artifacts introduced by the data integration technique. Several normalization strategies were carefully considered in combination with batch variance correction. The "Variance stabilization" technique was chosen due to its gentle normalization on individual microarray sets and because of both between color within array and between arrays systematic variances normalization (Huber et al., 2002). The missing biological variance recognition between the developmental stages may be the output of residual technical variance which was not removed after the data normalization. While Affymetrix technology is a single channel array with independent intensities for single microarray sample, the Agilent technology is strongly dependent on the experimental design. Differential gene expression in Agilent microarrays is usually computed by taking into account the experiment design. Samples are labeled with a different dye where for example the control sample with one color could be used as a direct reference, a reference with dye swap or a common reference for a group of other samples take into account the dye biases. The common data integration applied in the current study was lacking of individual experiment design and every microarray sample was counted as independent quantification. The experiment design could improve the quality the of data integration for future studies.

Given the robustness of the integrated Affymetrix microarrays, it was thus preferred to proceed with downstream data analysis with this dataset.

4.2 Machine learning and feature selection

The two microarrays data sets were used for distinguishing molecular biomarkers that are characterizing the morpholino knockdown treatment. Two feature selection strategies were used for identifying those genes that are informative for predicting unknown microarrays samples. Accuracy of the model was considered as an indicator of the feature selection performance.

The majority of the models showed high prediction accuracy differs from the random guess, which was used as an indicator for machine learning overfitting. The models that performed optimally were the models recognizing linear boundaries for the Affymetrix dataset. The Gaussian Linear model is based on the probabilistic distribution around a linear function of the data instances. The Support Vector Machine (SVM) on the other hand is a hyperplane separation based on non-probabilistic geometrical principles. The SVM showed to have identical prediction accuracy when compared to the Gaussian Linear model with C parameter set at 1. The use of the two models was in this case redundant since they both have had identical prediction performance for all the models. Glmnet, which is the Lasso and Elastic-Net Regularized Generalized Linear Models, was the second best performing algorithm based on penalized maximum-likelihood. The main strength of the algorithm is the exploration of a grid of tuning parameters and the elastic net penalty which cycles repeatedly until the optimal parameters converge. The performance for the quadratic boundaries was high for the tree-based algorithms while the Naïve Bayes algorithm showed low prediction accuracy. The SVM polynomial model, despites the high accuracy score, was the model that had the lowest distance from the scrambled control. The high scrambled performance indicates that models exploring polynomial boundaries need to be used carefully utilized and preferred when other models do not show a good prediction performance. The Partition Around Medoids (PAM), a variant of the k-means where the partition is performed around the means, was also used here for finding valuable data separation. The algorithm was however unable to reach a reliable accuracy probably because of the high dimensional data and the difficulty to assign a medoid center.

The Agilent data set was explored despites the challenges of integration the data sets for analysis. The best performing method was in this case the filtering feature selection methods which performed better than the embedded methods. The best performing method was the t-test coupled to the KNN model. The Relief filter feature selector showed improved prediction accuracy on high dimensionality data such as microarrays. In the case of Agilent, the Relief could be considered a valuable alternative to the t-test which however is computationally expensive and requires extensive optimization. The minimum redundancy maximum relevance feature selector was tested with two different quantity of seeds (1 or 10). The performance of the algorithm is greatly improved with a higher number of relevant features. This algorithm needs to be used however only in combination with other feature selectors for obtaining reliable starting seeds. In this case, the seeds were obtained from the t-test feature selector but other options could be considered for improving mRMR. This approach is not recommended due to the large amount of time invested to improve the prediction accuracy and long computational efforts.

4.3 Side-effects of morpholino antisense

The preliminary analysis carried out at the beginning of the study highlighted that the morpholino injection itself is not able to strongly deregulate gene expression. The PCA transformation is not able to specifically separate gene expression profiles by their treatment.

The most marked result from the data analysis is the increase of *phlda3* and *p53* genes which code for pro-apoptotic factors. *p53* is shown to positively regulate the expression of phlda3 in the apoptotic pathway thus probably suggesting the causative relation between the two (Kawase et al., 2009). According to the Gene Set Enrichment Analysis the *p53* activation is mainly due to the intrinsic apoptotic pathway. *p53* occur in response to the DNA damage which is likely to activate also cell death processes (Liu and Kulesz-Martin, 2002). Thermal melt results studying the binding affinity of morpholino to complementary biological and oligonucleotide molecules showed that the binding strength of morpholino to RNA is higher when comparing the binding to DNA. However, the difference in the melting curves of the two heteroduplex types is only of few °C thus showing a similar preference of morpholino for the two biological molecules (GeneTools, 2015). The high number of antisense molecules in the cell environment potentially binds to genomic DNA when single-stranded. Considering the morpholino

properties obtained *in vitro*, a possible mechanism that explains apoptosis activation in morpholino injected animals is the induction of DNA damage when the morpholino is targeted to a specific sequence. This mechanism is likely to occur during DNA replication where the two DNA strands are locally unzipped and the morpholino molecule could access. The eventual damage to DNA may upregulate the *p53* as direct response, inducing cell death through apoptotic pathways (Zhou and Prives, 2003). The pathway data analysis performed with Ingenuity Pathway Analysis confirmed the activation of the *p53* pathway and also disclosed the possibility of upregulation of the immune system response due to Interleukin-8 (IL-8). The latter is a neutrophil chemotactic factor which recruits the granulocytes to the site of infection. Its primary role is to induce localized infection reactions (Harada et al., 1994). In the case of morpholino, the IL-8 pathway is upregulated and thus, instead of infection, probably responding to cell death.

The stage and treatment interaction effect studied with two-way ANOVA and linear model mainly indicated the segmentation stage is altered after morpholino treatment. The increased gene expression observed in the segmentation stage confirms the off-target effects mainly given by neural death documented by Robu and colleagues (Robu et al., 2007). The latter study observes an altered zebrafish phenotype which is not due to the specific silencing of a target but due to cytotoxicity and *p53* activation. The stage specific gene upregulation observed during the segmentation stage was carefully analyzed using the information from the microarrays published research. The phenotypes that are observed after the morpholino injection involve different biological functions which show that there is no evident phenotypic bias towards one specific target gene function. The genes used for the study involve a heterogeneous group of phenotypes which makes the study robust against misleading dataset pooling. In order to know whether the stage specific upregulation of the genes shown in Figure 10 is an actual off-target effect or the result of the morphant phenotypes in the study, the information regarding the experimental conditions was again carefully analyzed. The RNA used for the microarrays at the segmentation stage was collected after observing the phenotype. Whether this gene upregulation could be considered as general effect needs however to be determined through a time course experiment comparison to mutant gene expression profiles.

ACKNOWLEDGEMENTS

I would like to thank Andrea Rossi and Christian Helker for the technical background and helpful information regarding morpholino antisense injection. I would like also to thank Soraya Hölper, Carol Yang, Chritian Groß and Robbin Bauwmeister for their support and precious suggestions regarding the data analysis and interpretation.

5 REFERENCES

Affymetrix, http://www.affymetrix.com, last access: May 2015

Agilent, http://www.agilent.com/home, last access: May 2015

Bill BR., Petzold A.M., Clark K.J., Schimmenti L.A., Ekker S.C. (2009) A primer for morpholino use in zebrafish. *Zebrafish*; **6(1)**: 69-77.

Baker K. and Parker R. (2004) *Nonsense-mediated mRNA decay: terminating erroneous gene expression*, Vol 16, Issue 3, 293–299.

Carole L., Yauk C. L., Berndt M. L, Williams A., Douglas G.R. (2004) Comprehensive comparison of six microarray technologies. *Nucleic Acids Res.* **32(15)**, e124.

Cermak T, Doyle EL, Christian M., Wang L., Zhang Y., Schmidt C. Baller JA., Somia NV, Bogdanove AJ and Voytas D. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting, *Nucleic Acids Res*.; **39(12)**, e82.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. (2013) Multiplex genome engineering using CRISPR/Cas systems, *Science*. **339(6121)**, 819-23

Davidson D.H (1986) *Gene activity in early development*, 3rd Edition, Academic Press, Inc.

Davis S., Meltzer P. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **14**, 1846–1847.

Ding C., Peng H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol.*,**3(2),** 185-205.

Draper B.W., Morcos P.A., Kimmel C.B. (2001) Inhibition of zebrafish fgf8 pre-mRNA splicing with morpholino oligos: a quantifiable method for gene knockdown. Genesis, **30(3)**, 154-6.

Edgar R., Domrachev M. and Lash AE. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30(1)**: 207–210.

Eisen JS and Smith JC. (2008) Controlling morpholino experiments: don't stop making antisense. *Development*. **135(10)**: 1735-43.

Eden E., Navon R., Steinfeld I., Lipson D., Yakhini Z. (2009) GOrilla: A Tool For Discovery And Visualization of Enriched GO Terms in Ranked Gene Lists. *BMC Bioinformatics*, **10**, 48.

Gabriel K. R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, ***58***, 453–467.

GeneTools, https://store.gene-tools.com/, last access: May 2015

GeneTools, http://www.gene-tools.com/history_production_and_properties, last access June 2015

Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/, last access: 19 March 2015

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*., **286**, 531–537

Harada A., Sekido N., Akahoshi T., Wada T., Mukaida N., Matsushima K. (1994) Essential involvement of interleukin-8 (IL-8) in acute inflammation. *J Leukoc Biol.*, **6(5)**, 559-64.

Huber W., von Heydebreck A., Sültmann H., Poustka A., Vingron M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*., **18**, Suppl 1, S96-104.

Hsu H.H, Lu M.D. (2008) Feature Selection for Cancer Classification on Microarray Expression Data, Eighth International Conference on Intelligent Systems Design and Applications

Jagga Z. and Gupta D. **(2014)** Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms, *BMC Proceedings*, **8**(Suppl 6): S2

Janson C. and During M. (2006) Morpholinos and Related Antisense Biomolecules, *Springer US, Springer-Verlag US*.

Johnson W.E., Li W., Meyer C.A., Gottardo R., Carroll J.S., Brown M., Liu X.S. (2006) Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A.*, **103(33)**, 12457-62

Johnson, WE, Rabinovic, A, Li, C (2007) Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics*, **8(1)**, 118-127

Kawase T., Ohki R., Shibata T., Tsutsumi S., Kamimura N., Inazawa J., Ohta T., Ichikawa H., Aburatani H., Tashiro F., Taya Y. (2009) PH domain-only protein PHLDA3 is a p53-regulated repressor of Akt. *Cell.*, **136(3)**, 535-50

Kok FO, Shin M, Ni CW, Gupta A, Grosse AS, van Impel A, Kirchmaier BC, Peterson-Maduro J, Kourkoulis G, Male I, DeSantis DF, Sheppard-Tindell S, Ebarasi L, Betsholtz C, Schulte-Merker S, Wolfe SA, Lawson ND. (2015) Reverse genetic screening reveals poor correlation between morpholino-induced and mutant phenotypes in zebrafish. *Dev Cell.*, **32(1)**, 97-108.

Lawson N.D. and Wolfe S.A. (2011) Forward and reverse genetic approaches for the analysis of vertebrate development in the zebrafish. *Dev Cell.*, **21(1)**: 48-64

Lee I., Lushington GH and Visvanathan M. (2011) A filter-based feature selection approach for identifying potential biomarkers for lung cancer, *Journal of Clinical Bioinformatics*, **1**, 11

Liang S., Li Y., Be X., Howes S., Liu W. (2006) Detecting and profiling tissue-selective genes. *Physiol Genomics*, **26(2)**, 158-62

Liu Y., Kulesz-Martin M. (2002) p53 protein at the hub of cellular DNA damage response pathways through sequence-specific and non-sequence-specific DNA binding. *Carcinogenesis*, **22 (6)**, 851-860.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. (2013) RNA-guided human genome engineering via Cas9. *Science.*, **339(6121)**, 823-6

Nasevicius A., Ekker S.C. (2000) Effective targeted gene 'knockdown' in zebrafish *Nat Genet*., **26(2)**, 216-20.

Pandey S. N., Lee Y., Yokota T., Chen Y. (2014) Morpholino Treatment Improves Muscle Function and Pathology of Pitx1 Transgenic Mice, *Mol. Ther*. **22 2**, 390–396.

Piccolo S.R., Sun Y, Campbell J.D., Lenburg M.E., Bild A.H., Johnson W.E. (2012) A single-sample microarray normalization method to facilitate personalized-medicine work- flows. *Genomics*, **100(6),** 337–344.

Pirooznia1 M., Jack Y Yang J. Y., Mary Qu Yang M. Q, Deng1 Y (2008) A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, **9(Suppl 1),** S13

http://www.ingenuity.com/products/ipa, last access June 2015

Rhodes DR., Barrette TR., Rubin MA, Ghosh D. and Chinnaiyan AM. (2002) Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer, *Cancer research*, **62**, 4427–4433

Robnik-Šikonja M, Kononenko I. (1997). An adaptation of Relief for attribute estimation in regression. *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97),* 296-304

Robu ME, Larson JD, Nasevicius A, Beiraghi S, Brenner C, Farber SA, Ekker SC. (2007) p53 activation by knockdown technologies. *PLoS Genet.*; **3(5)**: e78.

Saeys Y, Inza I, Larrañaga P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*., **23(19)**, 2507-17.

Schulte-Merker S. and Stainier D. (2014) Out with the old, in with the new: reassessing morpholino knockdowns in light of genome editing technology, *Development*, **141**, 3103-3104.

Smilde A.K, Jansen J.J, Hoefsloot H.C.J, Lamers R.J.A.N, Greef J.V.D, Timmerman M.E. (2005). ANOVA-Simultaneous Component Analysis (ASCA): A New Tool for Analysing Designed Metabolomics Data. *Bioinformatics*, **21(13)**, 3043–3048.

Smyth G.K., Ritchie M., Silver J., Wettenhall J., Thorne N., Langaas M., Ferkingstad E., Davy M., Pepin F., Choi D., McCarthy D., Wu D., Oshlack A., de Graaf C., Hu Y., Shi W., Phipson B. (2011). limma: Linear Models for Microarray Data. R package version 3.12.1

Streisinger G., Walker C., Dower N., Knauber D. and Singer F. (1981) Production of clones of homozygous diploid zebra fish (Brachydanio rerio). Nature, **291**: 293 - 296.

Summerton JE. (2007) Morpholino, siRNA, and S-DNA compared: impact of structure and mechanism of action on off-target effects and sequence specificity. - *Curr Top Med Chem*., **7(7)**: 651-60.

Summerton JE. and Weller D. (1997) Morpholino antisense oligomers: design, preparation, and properties. *Antisense Nucleic Acid Drug Dev.*, **7(3)**: 187-95.

Takei Y., Kadomatsu K., Yuasa K., Sato W., Muramatsu T. (2005) Morpholino antisense oligomer targeting human midkine: its application for cancer therapy. *Int J Cancer*., **114(3)**, 490-7.

The Gene Ontology Consortium, Ashburner M., Ball C. A., Blake A. J., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., David P. Hill, Issel-Tarver L., Kasarskis A., Lewis S., John C. Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., Sherlock G. (2000) Gene Ontology: tool forthe unification of biology. *Nat Genet.*, **25(1),** 25-9.

Ward A.J., Norrbom M., Chun S., Bennett C.F., Rigo F..(2014) Nonsense-mediated decay as a terminating mechanism for antisense oligonucleotides. *Nucleic Acids Res.*, **42(9)**, 5871-9

Yang H., Zhou Y., Gu J., Xie S., Xu Y., Zhu G., Wang L., Huang J., Ma H., Yao J. (2013) Deep mRNA sequencing analysis to capture the transcriptome landscape of zebrafish embryos and larvae PLoS One., **8(5)**, e64058

Zhou J., Prives C. (2003) Replication of damaged DNA in vitro is blocked by p53 Nucleic Acids Res., **31(14)**, 3881–3892.