Identifying a Diagnostic Classifier of CdLS by a Stepwise, Multi-platform Procedure

Zhe Zhang[1], Jinglan Liu[2], Dan Wilson[5], Mani Kaur[2], Matt Deardorff[2,4], Dinah Clark[2], Eric Rappaport[3], Michael Morrow[5], Ian Krantz[2,4]

[1]Center for Biomedical Informatics, [2]Division of Human Genetics, [3]Nucleic Acid/Protein Core Facility,
The Children's Hospital of Philadelphia,
[4]The University of Pennsylvania School of Medicine, [5]Xceed Molecular

Introduction

Cornelia de Lange Syndrome (CdLS) (OMIM 122470) is a dominant disorder of multiple congenital anomalies including characteristic facial features, upper limb defects, growth and cognitive retardation and other systemic abnormalities. Although mutations in the *NIPBL*, *SMC1A* or *SMC3* genes can be identified in about 60% of CdLS probands, molecular diagnosis is not available for the remaining patients. All three causative genes encode subunits or regulators of the Cohesin complex, indicating the existence of a general, Cohesin-related mechanism in CdLS. In this study, we performed a genome wide screen to identify a gene expression profile for CdLS under the hypothesis that defects in Cohesin complex may have a common effect at the gene expression level. We started the screen using comprehensive human genome microarrays (Affymetrix Inc.) and refined it using focused arrays (Xceed Molecular).

Affymetrix HG U133P2 Array

Training Samples

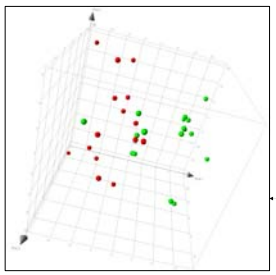
17 healthy controls vs. 16 severely affected CdLS patients

Balanced by age, gender, race (all Caucasians), etc.

54,675 Probe sets

Expressed in at least part of the training samples at sufficient level

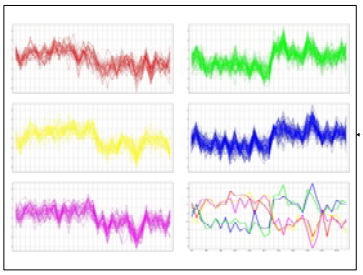
27,995 Probe sets



Principle Component Analysis (PCA) of 17 healthy controls (green) and 16 severely affected CdLS patients (red). The separation of the training groups indicates that they have different gene expression patterns.

Differentially expressed between controls and severe patients with FDR < 0.01

420 Probe sets



Differentially expressed genes were grouped by K-means clustering while K was set to 5 by the *gap* method (Tishirani et. al. 2000). Each vertical line represents one of the 33 training samples. The last plot shows the group average of 5 gene clusters, with matching colors. 4 to 9 genes were selected from each cluster (totally 32 genes) as focused genes for further validation.

Objectives:

1) Identify a gene expression profile as a potential diagnostic marker of CdLS using a stepwise approach
2) Explore the possibility of applying the same approach to other complex developmental disorders
3) Collect evidence about the causative effect of Cohesin defects on CdLS phenotypes at the gene expression level.

Comparison of Classification Methods Based on Cross-Validation

Method	TP	FP	FN	TN	Sensitivity	Specificity	Accuracy	Odds Ratio	Kappa
Diagonal Linear Discriminant Analysis	429	83	71	517	0.858	0.862	86.0%	37.427	0.718
Linear Discriminant Analysis	415	124	85	476	0.830	0.793	81.0%	18.680	0.619
Score for Expression Profile	430	82	70	518	0.860	0.863	86.2%	38.585	0.722
Naive Bayes	443	95	57	505	0.886	0.842	86.2%	41.063	0.723
Support Vector Machine	438	94	62	506	0.876	0.843	85.8%	37.817	0.716
K-Nearest Neighbor*	445	91	55	509	0.890	0.848	86.7%	44.973	0.734
Nearest Centroid	437	80	63	520	0.874	0.867	87.0%	44.817	0.739

TP: true positive; FP: false positive; FN: false negative; TN: true negative; Kappa: agreement between classification and diagnosis.
* The given results are based on K = 7, where the NN method has the best performance on this dataset

Nearest Centroid was selected as the optimal classification method, and was used through later stages of this study. The principle of this method is to classify a testing sample to the group whose centroid is closest. In this study, it was carried out with the following steps:

I. Select N genes differentially expressed between 2 training sample groups according to a certain statistical test.
II. Adjust the data of each selected gene in all samples by its pooled standard deviation (SDp) in training samples:
$X_{adj} = X / SD_p$
III. For each gene, calculate its median of X_{adj} in each training group, and generate an N-vector of medians as the centroid of each class.
IV. Measure the distance of each testing sample to both centroids as Pearson's correlation coefficients, r_0 and r_1.
V. Calculate a ***Discriminant Score (DS)*** as:
DS = 100*log2(corr2/corr1);
By default, a testing sample will be classified as a patient if DS > 0.

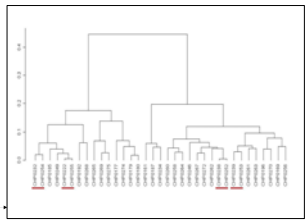
Xceed Focused Array

Sample Groups

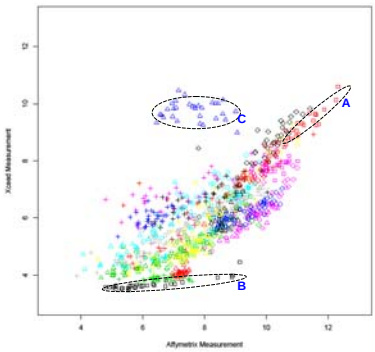
	Phenotype	NIPBL Mutation	#Samples
Original training (all Caucasians)	Healthy control	No	17
	Severe CdLS	Yes	14
New samples (with Caucasian background)	Healthy control	No	4
	Severe CdLS	Yes	6
	Moderate CdLS	Yes	9
	Mild CdLS	Yes	26*
	Moderate CdLS	No	4
	Mild CdLS	No, but SMC1A	9
	Mild CdLS	No	8
	Other disease	No	4
Non-Caucasians	Various	Various	8

*15 randomly selected samples of this group were used as training samples later.

32 Genes



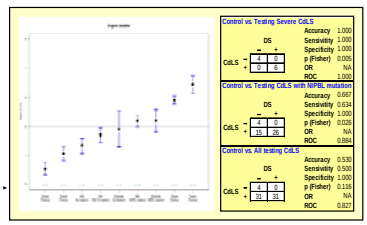
Clustering of 31 training samples based on Xceed data. Also included are 4 pairs of technical replicates. All replicates are nearest neighbor of each other, indicating that technical variance is smaller than biological variance.



The agreement of two platforms assessed by measurements of 32 focused genes in 31 training samples. Each color/shape combination represents a gene. While the data scale could be similar (Circle A) or very different (Circle B) between platforms, all genes have a positive Affymetrix-Xceed correlation with one exception (Circle C).

Differentially expressed between controls and 14 severe patients with p < 0.01

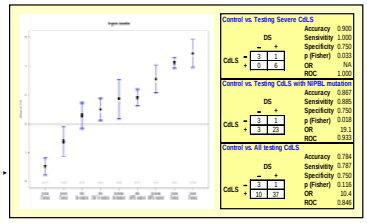
23 Genes



Evaluation of the 23-gene classifier by testing samples. The left figure shows the mean and standard error of Discriminant Scores of each sample subgroups. The right panel lists the results of classification while the range of testing samples is increased.

Differentially expressed between controls and 15 mild patients (with NIPBL mutation) with p < 0.01

10 Genes



Evaluation of the 10-gene classifier by testing samples. The left figure shows the mean and standard error of Discriminant Scores of each sample subgroups. The right panel lists the results of classification while the range of testing samples is increased.

Conclusion

In this study, we applied a well-controlled stepwise approach to obtain a gene expression profile for CdLS from two different array platforms. The measurements of gene expression were significantly correlated between platforms. Analysis of training samples identified 10 genes differentially expressed between healthy controls and CdLS patients with high confidence. A classifier based on the expression pattern of these genes correctly classified 40 of 51 (78.4%) testing samples that included multiple CdLS subtypes. Furthermore, quantitative representation of the classifier was significantly correlated to the severity of CdLS (p = 0.0001). The constructive outcomes of this approach encourages its potential application to other congenital disorders. Our results, in combination with findings of other independent research groups, also confirmed connection between Cohesin complex-regulated gene expression and CdLS phenotypes.