

# Definition of genomic regions potentially accessible to the transcriptional machinery by microarray expression profiling

Francesca Zolezzi

Axxam srl

San Raffaele Biomed. Science Park  
Via Olgettina, 58, 20132–Milan, Italy  
+39 02 2105681

francesca.zolezzi.fz@axxam.com

Tod A. Flak

Axxam srl

San Raffaele Biomed. Science Park  
Via Olgettina, 58, 20132–Milan, Italy  
+39 02 2105681

tod.flak.tf@axxam.com

Raffaele A. Calogero

University of Torino

Dept. of Clinical and Biol. Sciences  
Regione Gonzole 10, Orbassano, Italy  
+39 0116705410

raffale.calogero@unito.it

## ABSTRACT

Microarray analysis is frequently used to identify groups of genes co-regulated under specific treatments, which can be subsequently investigated to find common regulatory modules in their promoters. Although specific gene regulatory elements can be mapped in promoter regions, there are no bioinformatic tools that can infer whether these genes are or are not accessible to the transcription machinery. Thus, bioinformatic genome-wide approaches used to predict genes responding to specific stimuli (e.g. hormonal response, stress stimuli, etc.) lack information related to tissue specific accessibility of genomic loci. The identification of chromosomal regions that are potentially accessible to the transcriptional machinery could therefore be useful in improving the quality of prediction of gene expression. As the initial step in this study we analyzed microarray expression data from multiple tissues to define chromosomal regions which seem to exhibit reduced transcriptional level across multiple genomic loci.

## Categories and Subject Descriptors

J.3 [Life and Medical Sciences]: *Biology and genetics.*

## General Terms

Design, Experimentation.

## Keywords

Microarrays, human tissues, chromosomes.

## 1. INTRODUCTION

The basis of eukaryotic complexity and phenotypic variation lies in a highly controlled architecture composed of distinct systems:

transcriptional expression modulation, post-transcriptional mRNA stability, translational expression modulation and post-translational protein modification/stability. Of particular interest to the drug discovery field are the issues related to “in silico” production of models of transcriptional regulatory networks and their relations to human pathologies. In fact, a better understanding of transcriptional regulatory networks might facilitate the identification of key components that can be used as targets for drug discovery. Although computational approaches can perform the mapping of transcriptional regulative elements (TRE) and the identification of transcription modules (i.e. groups of TRE acting together to induce a specific transcriptional response), there are no bioinformatic tools able to infer if these loci are accessible to the transcription machinery. Therefore the identification of chromosomal regions that are transcriptionally active, and hence accessible to the transcriptional machinery, could improve the quality of prediction of regulatory modules affecting gene transcription upon specific external stimuli.

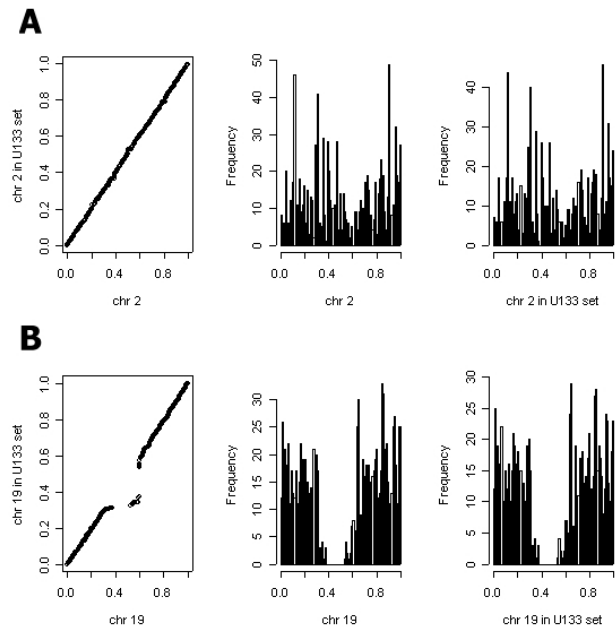
## 2. MATERIAL AND METHODS

Tissue samples were obtained through EC-approved collaboration arrangements with clinical departments, strictly following national and international guidelines for handling tissue specimens and safeguarding patients’ clinical and personal data. Total RNA was extracted using standard methods. RNA quality and integrity was assayed using the Bioanalyser 2100 (Agilent). Only samples characterized by low degradation and sharp 18S, 28S peaks were used for microarray analysis. cRNA probes were prepared as suggested by Affymetrix. Hybridization and scanning of Affymetrix HG-U133 GeneChip<sup>®</sup> array set (A and B) were performed using standard conditions suggested by the manufacturer. Probeset expression analysis was performed with Affymetrix MAS 5.0 software [1]. Data annotation and handling was mainly done with Bioconductor [2]. To summarize the MAS expression calls, we performed two steps. First, for each probeset and tissue type, we computed the ratio of P (present) calls to the number of replicates for that tissue type. Then we computed the average of this value for all probesets related to the same LocusLink record, to arrive at the expression index (EI) for a gene, having a range between 0 and 1. An indicator called GT (Generally Transcribed) was defined as true for genes having an EI>0.35 in all tissues analyzed.

### 3. RESULTS

#### 3.1 Chromosomal distribution of probesets

We observed that the 44928 probesets of the Affymetrix HG-U133 GeneChip® array set (A and B) were associated to a total of 18878 LocusLink records (<http://www.ncbi.nih.gov/locuslink>). Using the Bioconductor humanCHRLOC (version 1.4.1) data package we obtained the absolute chromosomal location of the transcriptional start site for 15562 of these LocusLink identifiers. To evaluate if the chromosomal distribution of the probesets present in the arrays is similar to the chromosomal distribution of the mapped human genes, we normalized the positions of the probesets and of the human genes in relation to the full length of the chromosomes and compared them by a quantile-quantile plot. For almost all of the chromosomes it was observed that the array probesets are distributed on the human chromosomes in a way similar to the mapped genes (Figure 1A). The probesets showed a different distribution from the genes only on chromosome 19, around the centromer where the genes are underrepresented in the arrays (Figure 1B).



**Figure 1: Overlap between chromosomal data and U133 set.**

These results indicate that expression profiling based on the HG-U133 set can be used to identify chromosomal regions characterized by tissuespecific transcription factor chromatin accessibility.

#### 3.2 Tissue for transcription profiling

We started a broad microarray-based transcription profiling study of normal human tissues. To this point we have profiled 11 different tissues and 2 cell lines; six additional tissues are in the pipeline (Table 1). For most tissues, at least three experiments were performed on samples derived from different individuals.

**Table 1. Tissues for transcription profiling**

Tissue Type	Profiled	Replicates
Liver	Y	3
Kidney Cortex	Y	3
Kidney Medulla	Y	3
Myocard	Y	3
Vein	Y	2
Aorta	Y	5
Bone	Y	3
Bone Marrow	N	3
Adipose Tissue	Y	4
Ovary	Y	5
Bladder	Y	2
Prostate	Y	3
Hek	Y	3
HepG2	Y	3
Pancreas	N	3
Cortex gray matter	N	3
Dorsal Root Ganglia	N	3
Dorsal Horn	N	3
Ventral Horn	N	3

#### 3.3 Transcriptionally-active chromosomal regions

We wanted to explore whether microarray expression information could allow us to define chromosomal areas, which include several genes, that are transcriptionally active. This could permit us to infer if a gene, which is not expressed in a certain tissue or condition, has the potential to be expressed based upon its location within a transcriptionally-active area of the chromosome. A difficult issue in this analysis is that there is no absolute criteria to define whether a gene is expressed or not. We decided, at least at this preliminary phase, to start with an analysis based on “Present” calls from the Affymetrix MAS 5.0 software[1]. MAS expression calls (P: present, M: marginal, A: absent) are a measure of the difference between probe specific signal and unspecific hybridization. The expression index (EI) (see Materials and Methods) indicates how robust is the “presence” call for a specific gene. We then created an indicator we refer to as GT (Generally Transcribed), which simply categorizes each gene based upon having an  $EI > 0.35$  in all tissues analyzed. We identified 4702 genes that met this GT criterion (“GT genes”). Chromosome Y exhibited the lowest proportion of genes meeting this criterion (13.9%), while chromosome 13 showed the highest proportion (41.4%). By plotting the chromosomal positions of the GT genes, it is possible to define relatively large chromosomal regions characterized by tissue-dependent expression.. As an example, Figure 2 shows the frequency distribution of GT genes for chromosome 5. The arrow indicates a chromosomal region that lacks GT genes, and that might be associated to tissue-specific expression. The inset shows the hierarchical clustering of

