Alma mater studiorum - Università di Bologna

Dottorato in Biotecnologie Cellulari e Molecolari: XXIII ciclo
Settore scientifico disciplinare di a□erenza: BIO11

BIOINFORMATIC METHODS IN APPLIED GENOMIC RESEARCH

Presentato da
Ra□aele Fronza

Coordinatore Dottorato:
Prof. Santi Mario Spampinato

Supervisore:
Prof. Rita Casadio

Esame finale anno 2011

Abstract

Here I will focus on three main topics that best address and include the projects I have been working in during my three year PhD period that I have spent in different research laboratories addressing both computationally and practically important problems all related to modern molecular genomics.

The first topic is the use of livestock species (pigs) as a model of obesity, a complex human dysfunction. My efforts here concern the detection and annotation of Single Nucleotide Polymorphisms. I developed a pipeline for mining human and porcine sequences. Starting from a set of human genes related with obesity the platform returns a list of annotated porcine SNPs extracted from a new set of potential obesity-genes. 565 of these SNPs were analyzed on an Illumina chip to test the involvement in obesity on a population composed by more than 500 pigs. Results will be discussed. All the computational analysis and experiments were done in collaboration with the Biocomputing group and Dr.Luca Fontanesi, respectively, under the direction of prof. Rita Casadio at the Bologna University, Italy.

The second topic concerns developing a methodology, based on Factor Analysis, to simultaneously mine information from different levels of biological organization. With specific test cases we develop models of the complexity of the mRNA-miRNA molecular interaction in brain tumors measured indirectly by microarray and quantitative PCR. This work was done under the supervision of Prof. Christine Nardini, at the "CAS-MPG Partner Institute for Computational Biology" of Shangai, China (co-founded by the Max Planck Society and the Chinese Academy of Sciences jointly)

The third topic concerns the development of a new method to overcome the variety of PCR technologies routinely adopted to characterize unknown flanking DNA regions of a viral integration locus of the human genome after clinical gene therapy. This new method is entirely based on next generation sequencing and it reduces the time required to detect insertion sites, decreasing the complexity of the procedure. This work was done in collaboration with the group of Dr. Manfred Schmidt at the Nationales Centrum für Tumorerkrankungen (Heidelberg, Germany) supervised by Dr. Annette Deichmann and Dr. Ali Nowrouzi.

Furthermore I add as an Appendix the description of a R package for gene network reconstruction that I helped to develop for scientific usage (`http://www.bioconductor.org/help/bioc-views/release/bioc/html/BUS.html`).

Publications (during PhD period)

Some of the results presented and discussed herein are contained in the following publications:

[1] BARTOLI, L., MONTANUCCI, L., FRONZA, R., MARTELLI, P. L., FARISELLI, P., CAROTA, L., DONVITO, G., MAGGI, G. P., AND CASADIO, R. The bologna annotation resource: a non hierarchical method for the functional and structural annotation of protein sequences relying on a comparative large-scale genome analysis. *J Proteome Res 8*, 9 (Sep 2009), 4362–71.

[2] BIASCO, L., BARICORDI, C., BARTHOLOMAE, C., BRIGIDA, I., SCARAMUZZA, S., FRONZA, R., MERELLA, S., AMBROSI, A., PELLIN, D., DI SERIO, C., MONTINI, E., VON KALLE, C., SCHMIDT, M., AND AIUTI, A. Uncovering haemopoietic system dynamics and tracking fate and long-term survival of single progenitor clones in vivo in gt patients by retroviral tagging. *American Society of Gene and Cell Therapy* (2011). Submitted.

[3] FONTANESI, L., FRONZA, R., SCOTTI, E., GALIMBERTI, G., CALO, D. G., BONORA, E., VARGIOLU, M., COLOMBO, M., CASADIO, R., ROMEO, G., AND RUSSO, V. The fagenomich project: a whole candidate gene approach to identify markers associated with fatness traits in pigs and investigate the pig as a model for human obesity. In *Second European Conference on Pig Genomics, Biotechnical Faculty, University of Ljubljana* (2008).

[4] FONTANESI, L., SCOTTI, E., GALIMBERTI, G., CALO, D. G., FRONZA, R., COLOMBO, M., MARTELLI, P. L., BUTTAZZONI, L., CASADIO, R., AND RUSSO, V. Selective genotyping analysis of 677 snps to identify markers associated with back fat thickness in italian large white pigs. In *Proceedings of the Second European Conference on Pig Genomics* (2009), pp. 978–961.

[5] FRONZA, R. An integrated bioinformatic platform for snps detection. short communication. In *Sixth Annual Meeting Bioinformatics Italian Society* (2009).

[6] FRONZA, R., FONTANESI, L., MARTELLI, P., RUSSO, V., AND CASADIO, R. Annotation of illumina's porcinesnp60 beadchip single nucleotide polymorphisms. *Animal Genetics* (2011). Submitted.

[7] FRONZA, R., FONTANESI, L., RUSSO, V., AND CASADIO, R. An integrated bioinformatic platform to query the pig genome. In *ASPA, XVIII Congresso Nazionale* (2009).

[8] FRONZA, R., TRAMONTI, M., ATCHLEY, R. W., AND NARDINI, C. Joint analysis of transcriptional and post-transcriptional brain tumor data: Searching for emergent properties of cellular systems. *BMC Bioinformatics* (2011). In Press.

Contents

List of Figures

List of Tables

Part I

General Concepts

Chapter 1

Introduction

1.1 Obesity

Obesity is a pandemic, complex and multifactorial disease. It predispose to morbidity as type 2 diabetes, heart disease and hypertension with the consequence of premature mortality. Obesity is common all over the world where an *obesogenic* environment is created and where sedentary lifestyle and high energy intake are encouraged. Although this trend of increasing body fat is driven by the *obesogenic* environment, it is facilitated by the individual's genetic susceptibility to excessive weight gain.

This disorder is highly heritable and studies in twins, adoptees and families shows that 80% of the variance in the body mass index (BMI) is attributable to genetic factors. Relative risk of obesity among siblings was estimated to be 3 to 7 [AFN96], the concordance rate of obesity is higher between monozygotic twins than dizygotic twins [MNE97, SFH86, AKK+96], and adoptees' weight is often closer to their biological parents than their adoptive parents [SSH+86]. These studies supported the role of genes in the pathogenesis of human obesity.

Obesity gene research has advanced rapidly over the past 2 decades, which has provided revelation of the molecular mechanism of energy homeostasis in the process. Traditional methods employed to uncover these obesity genes include genome-wide scans which studied unrelated obese individuals, linkage studies which examined related pairs or families with obesity, and association studies which investigated the association between obesity and polymorphic variants of candidate genes predicted to affect weight regulation. Differently from how they worked for other multifactorial disorders, these approaches have not been as promising for common obesity, because the obese phenotype is very heterogeneous, even within the same family.

The progress in computers and technologies such as high throughput sequencing machines have made it possible to analyse combinations of multiple SNPs or haplotypes in candidate genes related to weight regulation. A new approach to identify genes affecting weight regulation is to study individuals who are thin and have difficulty gaining weight (obesity resistance) [LKM+83, FO06] because like obesity, leanness is heritable and it is therefore conceivable that studies of the reverse phenotype can complement the efforts to uncover the obesity genes.

1.1.1 Obesity and livestock genomes

The human genome sequencing project had high impact on the animal genetic research and the complete genome sequences of several livestock species are now available. A basic underlying hypothesis of comparative genomics is that evolutionarily conserved sequences are functionally important. Based on this hypothesis, biologic experimentation has focused on sequences that are highly conserved among vertebrate species. Clues in elucidating molecular mechanisms and pathways involved in disorders can be obtained with animal models, in which genetic experimental manipulation is allowed [BGK08]. Depending on the disorder type, non-rodent animal models have been described as the better ones. The pig (*Sus scrofa*) in relation with

dimensions, physiology and genetics is considered a better in vivo model for obesity than other animal models like mouse and rat [SG08, BZ09]. Genetic maps of livestock genomes have been applied in several linkage studies to map loci and genes that underlie genetic variance of economically important traits. In pig production, fatness is an important trait that has been extensively studied over the last years.

1.1.2 Fatness regulation traits in pigs

Pigs have been intensively used as a biomedical research model for various human physiological conditions [TS96], and pigs and humans share numerous physiological and phenotypic similarities for fat deposition and food intake [HHP79]. Therefore, identified chromosomal regions and genes that regulate lean growth and fat deposition in pigs might be helpful to study the genetic basis of human obesity and other related health problems.

In pig production the two depots of highest interest are back fat (which accounts for 60-70% of total body fat) and intramuscular fat (IMF). In breeding programs, selection is directed against back fat thickness (BFT), because reduced carcass fatness benefits carcass quality and production efficiency. Over the years breeding has resulted in reduced back fat thickness (from 3.2 to 1.9 mm) and increased loin muscle area (40 to 60 cm^2) which indicates high genetic determination of these body composition traits [RPK03]. With the use of molecular genetic methods, such as candidate gene and genome scan approaches, the identification of genes for obesity and growth can be obtained. In *Sus scrofa*, candidate genes associated with obesity and growth include Leptin Receptor, Melanocortin-4 Receptor, Agouti related protein, Heart fatty acid binding protein 3, and Insulin-like growth factor 2. Some of these candidate genes also explain the variation in obesity levels in humans [RPK03].

In human obesity body fat is accumulated in four different depots in the body: subcutaneous, visceral, intermuscular and intramuscular [CCT99]. Identification of genes involved in fat development in pigs could lead to a better understanding of the underlying genetic mechanisms in pigs and, by projection, in humans.

1.1.3 Quantitative trait loci for complex traits

The candidate gene approach applies previous knowledge about the functions of genes to select those genes that might be involved in the trait of interest. The selected candidate genes are tested for association with the trait or phenotype. The candidate gene approach has particularly been successful for relatively simple traits with only few genes involved, like the identification of genes involved in coat color in pigs [KWT+98, KMPA01]. In the whole genome scan approach, a large number of genetic markers from across the genome is analysed to observe segregation of chromosomal segments through a pedigree. By analyzing the (co-)segregation of the phenotypic trait or value with certain chromosomal segments in an experimental population, chromosomal regions harboring genes affecting a quantitative trait can be identified ([Gel96]. Such a chromosomal region harboring genes that influence the variation of a quantitative trait is referred to as a quantitative trait locus (QTL). In general, experimental populations for QTL mapping are comprised of intercrosses of phenotypically divergent founder populations. The assumption that the genes influencing the trait of interest are fixed in these founder populations enables identification of QTL regions explaining the genetic differences among populations. For most detected QTL, the identified chromosomal regions are still large and harbor hundreds of genes. To narrow down these regions to the gene(s) or eventually the mutation causing the effect on the phenotype, fine-mapping of this region is necessary. Fine-mapping can be achieved by adding markers in the region, and/or by applying additional approaches like linkage disequilibrium mapping (LD) and haplotype-sharing analysis.

	C	T	G
A	*1*	*1*	*2*
C	*2*	*1*	
T	*1*		

Table 1.1. SNPs categories: transitions (*2*) and transversion (*1*)

1.1.4 Single nucleotide polymorphisms

According to Brookes [Gel96, Bro99], the definition of a single nucleotide polymorphism is a single base pair position in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), and where the less frequent allele has an abundance of 1% or greater. This definition is rather strict, as it holds some factors that exclude polymorphisms based on the allele frequencies in the population studied. In addition, insertion/deletion polymorphisms (indels) are excluded, since they are created through a different mechanism and do not have an alternative nucleotide on the polymorphic position. In practice, all subtle sequence differences, including indels, are potential markers for mapping purposes, even those with low frequencies. In addition, the frequency is dependent on the studied population. Consequently, a polymorphism might be a true SNP in one population and just a subtle DNA difference in another. Altogether, it is neither practical nor useful to classify all subtle DNA differences as either true SNPs or indels or 'low' frequency polymorphisms. Most polymorphisms can be used for the same goal, using the same methods, regardless of allele frequencies, population or origin of the mutation. For these reasons, in this thesis a wide definition of SNPs is applied and all point mutations are considered SNPs, as is commonly done in practice. SNPs are predominantly bi-allelic, even though in principle any of the four nucleotides can be present at any position in a stretch of DNA. This is due to the low frequency of mutations that leads to new SNPs. This mutation rate is estimated at 10^{-8} changes per nucleotide per generation and corresponds to about 100 new SNPs per individual. As a result, the chance for a second independent mutation that introduces a third allele at the same position is very low (10^{-16}, about 1 second independent mutation per 10^6 individuals).

Basically, the bi-allelic SNPs are comprised of two different categories: transitions and transversions (see table 1.1). In transitions, a purine is exchanged for the other purine, while on the reverse strand a pyrimidine is exchanged for the other pyrimidine. Transversions consist of purine-pyrimidine (and their complementary pyrimidine-purine) exchanges. The occurrences of transitions and transversions are not equal in the genome. Although there are two times as many options for transversions as for transitions, transitions are found twice as often as transversions. The higher level of transitions might be related to 5- methylcytosine (5mC) deamination reactions. 5mC is the product of post-synthetic modification of cytosine residues and occurs primarily in CpG doublets in the mammalian genome. 5mC is a mutable site that can undergo spontaneous deamination to thymine. Although a repair mechanism specifically recognizes G-T mismatches and replaces the thymine with cytosine, the 5mC to T transition mutation occurs about 10 times more often than other transitions [HG93].

The rate at which nucleotide differences are observed between two randomly chosen chromosomes is called the *nucleotide diversity index* (Nei1979). In the human genome, this index is estimated at one in a thousand basepairs [LS91]. Screening more chromosomes (more individuals) will identify more polymorphisms, but the nucleotide diversity index remains constant and allows comparisons among studies applying different sample sizes. Compared to the total human genome, the nucleotide diversity index in coding regions is about 4-fold lower and about half of these coding mutations result in non-synonymous codon changes [LS91]. Nucleotide diversity indexes are reported to be 1/1331 bp in humans [SWS+01], 1/515 bp in mice [LTWD+00] and 1/609 bp in pigs [FFS+02]. Overall, this adds up to about several millions of SNPs between any two individuals and about 100,000 amino acid changes in the proteomes. It has been estimated that in the world's human population, about 10 million SNPs (that is, 1

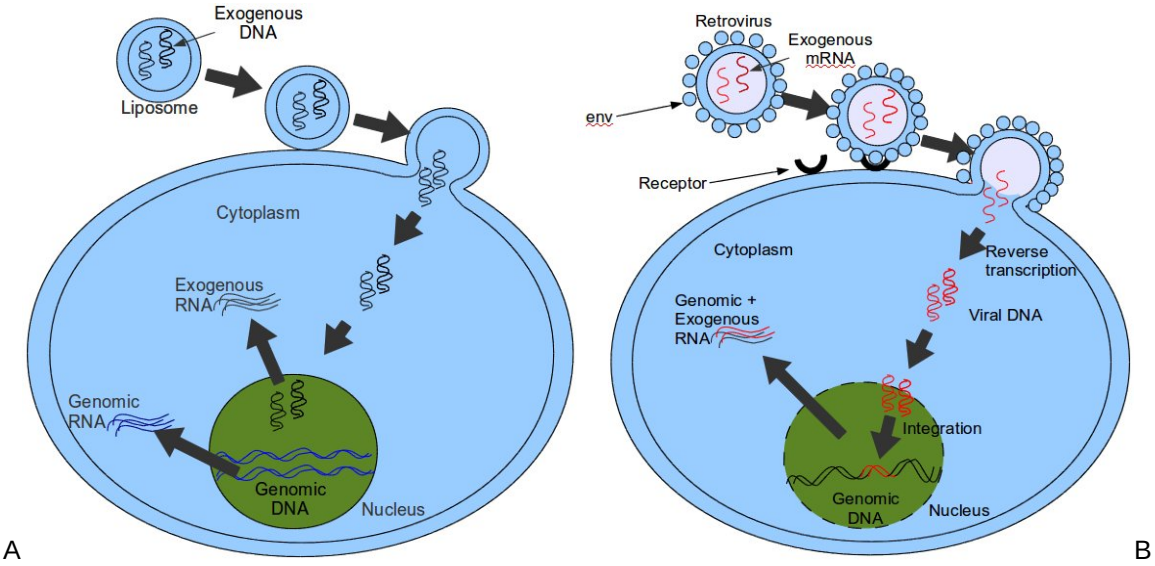


Fig. 1.1. Delivery system. A: schematic view of the delivery process mediated by liposomes; the exogenous material provided by the liposome enters (transiently) in the nucleus as episomic DNA. B: delivery process provided by retrovirus;

variant site per 300 bases on average) vary such that both alleles are observed at a frequency of 1% or greater, and that these 10 million common SNPs constitute 90 % of the variation in the population [RGA03]. I will use this last SNP frequencies estimation as a parameter to discriminate SNPs that arise from noise (paralogues, pseudogenes).

1.2 Insertion sites in gene therapy

1.2.1 Gene therapy

The major goal of gene therapy is to introduce a functional gene into a target cell and restore protein production that is absent or deficient due to a genetic disorder. Gene transfer has been successfully used to treat a wide spectra of genetic diseases like X-linked severe-combined immunodeficiency (SCID-X1), ß-thalassemia, adenosine deaminase deficiency, chronic granulomatous disease and more recently adrenoleukodystrophy [CHBAB+09].

This is a technique whereby the loss of function due to a faulty gene is restored by a working gene introduced in the host using a gene transfer strategy.

1.2.2 Gene transfer strategies

Over the years two categories of gene transfer vehicles have been developed: synthetic and virus-based gene delivery systems (see Figure 1.1) .

Synthetic system Synthetic gene delivery systems depend on direct delivery of genetic information into a target cell and include: a) direct injection of naked DNA and b) encapsulation of DNA with cationic lipids (Figure 1.2). These delivery systems exhibit low toxicity but gene transfer is in general ine□cient and often transient. For those reasons this group of delivery systems is employed to the delivery of drugs or to transiently transfect with DNA or RNA naked cells.

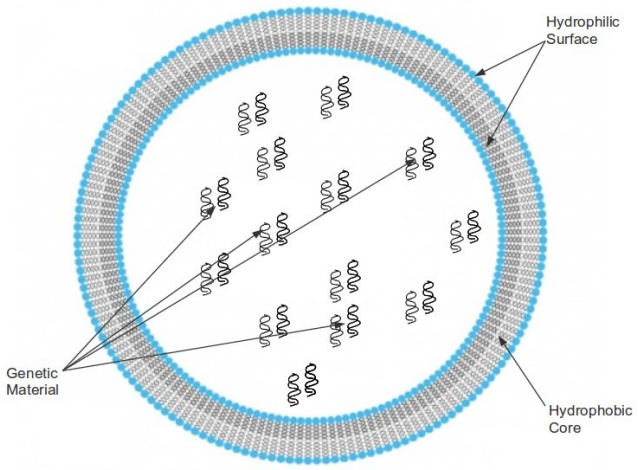


Fig. 1.2. A liposome

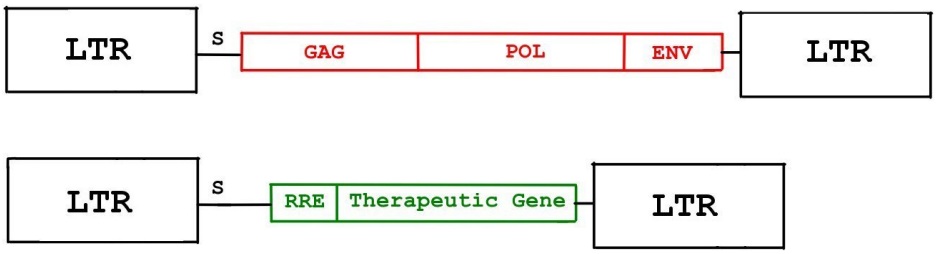


Fig. 1.3. Nonreplicating virus particles containing the genetic information of a therapeutic gene. GAG, POL and ENV viral structural genes; S, Packaging signal; LTR, Long Terminal Repeats; RRE, rev responsive element.

Virus-based system Viral delivery systems are based on replicating viruses that have the ability to deliver genetic information into the host cell. As shown in Figure 1.3in general, genomes of replicating viruses contain two kinds of sequences: 1) coding regions and *2) cis*-acting regulatory elements. The first sequences enclose the genetic information of the viral structural and regulatory proteins required for propagation of infectious viruses, whereas *cis*-acting sequences are essential for packaging of viral genomes and integration into the host cell. In a replication-defective viral vector, the coding regions of the virus are not functional and could be deleted and replaced by the genetic information of a therapeutic gene, leaving the cis-acting sequences intact (figure 1.3).

Viral vectors currently available for gene therapy are based on di□erent viruses and can roughly be categorized into: 1) vectors that have the ability to integrate their viral genome into the chromosomal DNA of the host cell (adeno-associated virus and retroviruses); 2) vectors that deliver their genomes into the nucleus of the host cell, remaining episomal (adenovirus and herpes simplex virus type 1).

1.2.3 Insertion sites

While most if the delivery systems adopted in gene therapy are based on integrating viruses, the positive clinical outcomes are balanced by events that could drive the patients to adverse consequences. It has been showed that the integration of the viral genome could increase the

transcription of cancer-related genes and the probability of developing leukemia [HBAVKS+03a, HBAvKS+03b].

Due to the adverse integration event, a strong interest has been aroused in order to follow the fate of the gene-corrected cells, monitor eventual adverse consequences and aid researchers and physicians to develop new strategies in gene therapy. The analysis of the insertional profile of the vectors is a tool used to verify the "safety" of the gene therapy approach in patients.

1.2.4 LAM-PCR

Different technologies are developed to detect the flanking regions of viral insertion sites in which polymerase chain reaction (PCR) is the base: LM-PCR [MW89], LAM-PCR [SSB+07] and nrLAM-PCR [GEP+09]. The last two methods have the advantage to be sufficiently sensitive, specific and robust. Besides the analysis of the insertions sites could be conducted on complex samples with multiple integration sites and starting from a minimal amount of material. On the contrary some disadvantages are affecting such techniques like:

- bias due to the use of restriction enzymes [WBM+10] where some sites could remain undiscovered
- the complexity of the techniques in which several steps are required to achieve the results.

1.3 Next generation sequencing

1.3.1 Sequencing technologies

1.3.1.1 Maxam-Gilbert sequencing It was the first powerful method developed in 1977 [MG77] for sequencing DNA molecules. It is based on a chemical digestion where a DNA chain is digested by chemical reagents that cut uniformly at the level of one nucleotide. In this method the DNA chains are marked at the 5' end by ^{32}P using a polinucleotide kinase. See figure 1.4 for details. This method was replaced by the chain-termination method.

1.3.1.2 Sanger sequencing The chain-termination method was published by Frederick Sanger in the same year [SNC77]. This methods is based on the controlled interruption of the enzymatic replication of the DNA. The DNA pol I is used for copying a single strand DNA chain starting from complementary fragments. If in the incubation mix are present, in addition to the four standard deoxynucleotides (dATP, dGTP, dCTP and dTTP), *four 2',3' dideoxy* analogs, the incorporation of an analog stops the synthesis with the result of the production of 4 populations of fragments having different length. If the four deoxynucleotides are marked radioactively an auto-radiography of an electrophoresis run could reveal the chain DNA sequence (see figure 1.5). Technical modifications in the procedure with the introduction of fluorescent terminator dyes, the use of optical systems and the availability of computers explain why Sanger method became at the end of the '80s the standard procedure in sequencing projects.

1.3.1.3 High-throughput sequencing After the completion of the human genome sequencing in 2003 the industry released sequencing technologies that produce millions of sequences in one run at a fraction of the cost necessary in the pre-high-throughput "era". Here I will describe the most used.

454 sequencing This system is based on pyrosequencing [RKP+96], in which single strand DNA chains are immobilized onto beads and amplified by PCR in order to have a clonal colony per bead. Then beads are placed into a small camera (~29μm) in the PicoTiterPlate via controlled process that attempts to deposit one bead per well. In the weel the pyrosequencing process is made adding luciferase and other components that generate light when a nucleotide

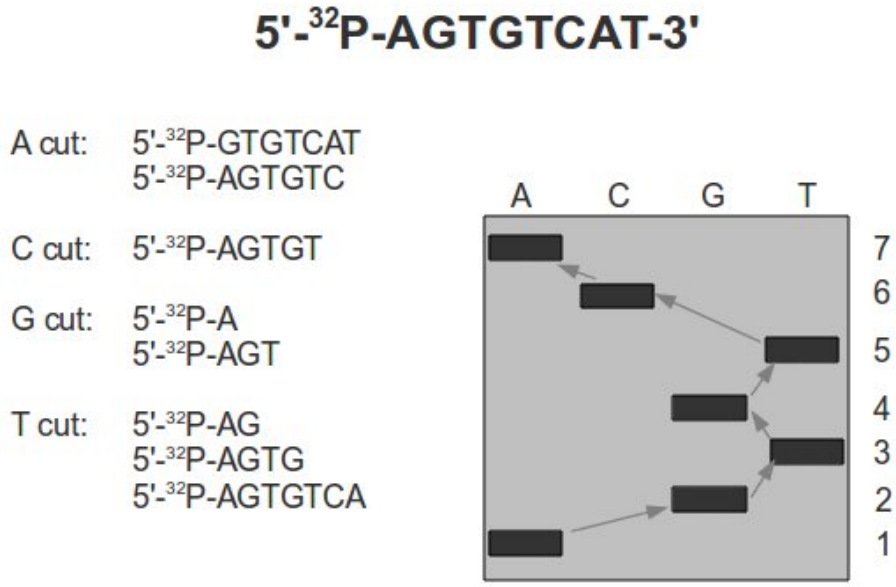


Fig. 1.4. Schematic diagram of the Maxam-Gilbert procedure. On the right the gel with the radioactive fragments after the specific cut of the sequence AGTGTCAT in each bases (A, C, G and T). Following the gray arrows the sequence is obtained.

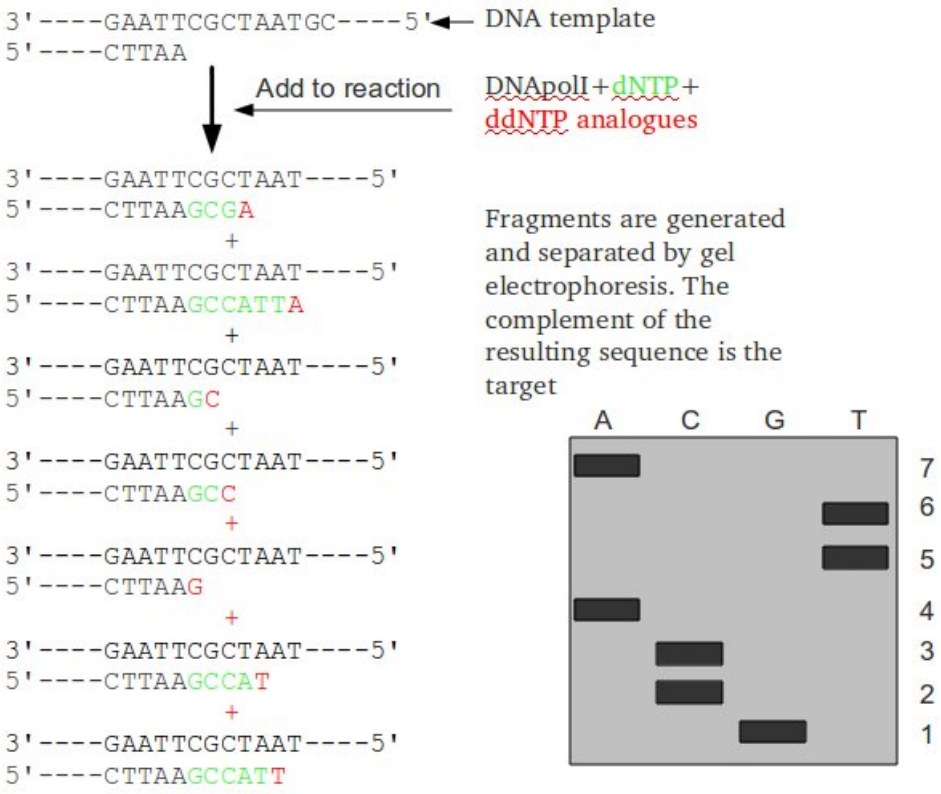


Fig. 1.5. Sanger procedure. Fragments are produced adding analogues of dideoxy nucleotides (red) in the polymerization mix. In the example the reaction returns 7 different fragments.

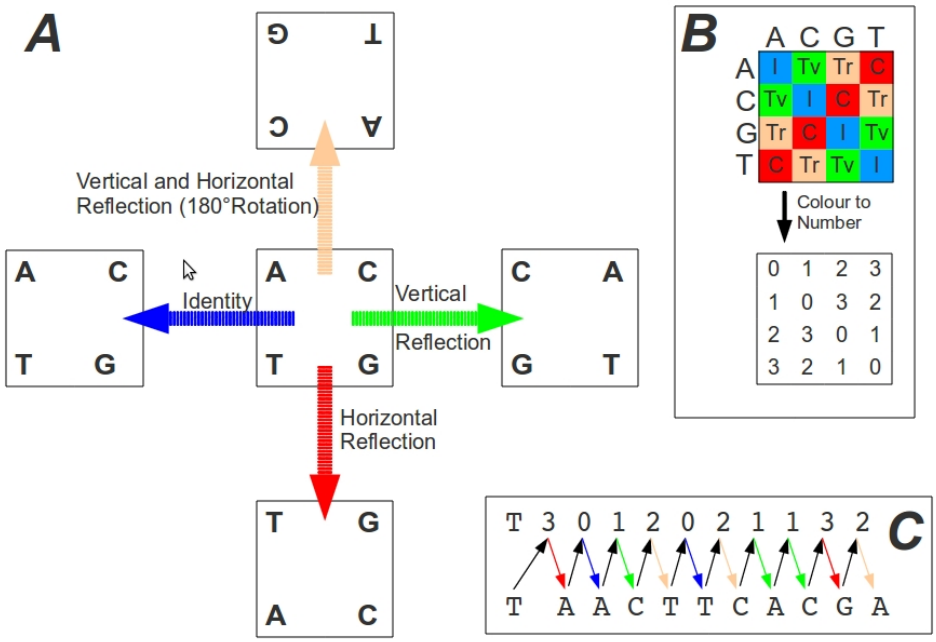


Fig. 1.6. Color space scheme in SOLID system.

is incorporated during the synthesis. The light is transferred via optic fiber (embedded in the PicoTiterPlate) to a CCD camera controlled by the software and that stores data in a binary proprietary format called Standard Flow-gram Format (SFF). This technology is able to sequence more than 1 million reads with more than 400 bases each.

Illumina (Solexa) sequencing The Illumina sequencing (previously called Solexa) is a method that uses reversible dye terminators. The idea is to attach the DNA molecules to synthetic primers on a chip and then to amplify locally the fragments in order to obtain a local clonal colony. The sequencing is done by cycles and for each cycle modified ddNTPs are added. A CCD takes the images of the fluorescent labeled nucleotides and the blocker at the 3' end of the dye is chemically removed from DNA. Performances of this system are hundred millions of 100 bases length reads in a week.

SOLiD sequencing This is the last technology released in the market (2007) and it is based on a new sequencing technology called (by SOLID) 2 Base Encoding. As for the other two technologies the DNA chains are isolated and then amplified locally in a weel. Then the SOLiD procedure is based on sequencing by ligation of specific 8-mer probes in which only the first two bases (starting at the 3' end) are complementary to the nucleotides being sequenced whereas the other 6 are degeneres. The procedure is based on an "external" cycle composed by m internal reactions. Refer to figure for details. I want to underline that in this procedure the fluorescence is related to the composition of a dinucleotide and that each base is interrogated two times. This points to a "new" sequence space called *color space* where each sequence expressed in this code needs to be decoded. The decoding step requires 1) to know which is the first base of the sequence and 2) a table that associate each color to a dinucleotide. The color scheme is the symmetry group of a rectangle (Klein four-group; Figure 1.6). It is interesting to point out that this code is useful to discriminate between sequence variation (SNPs) and sequencing errors.

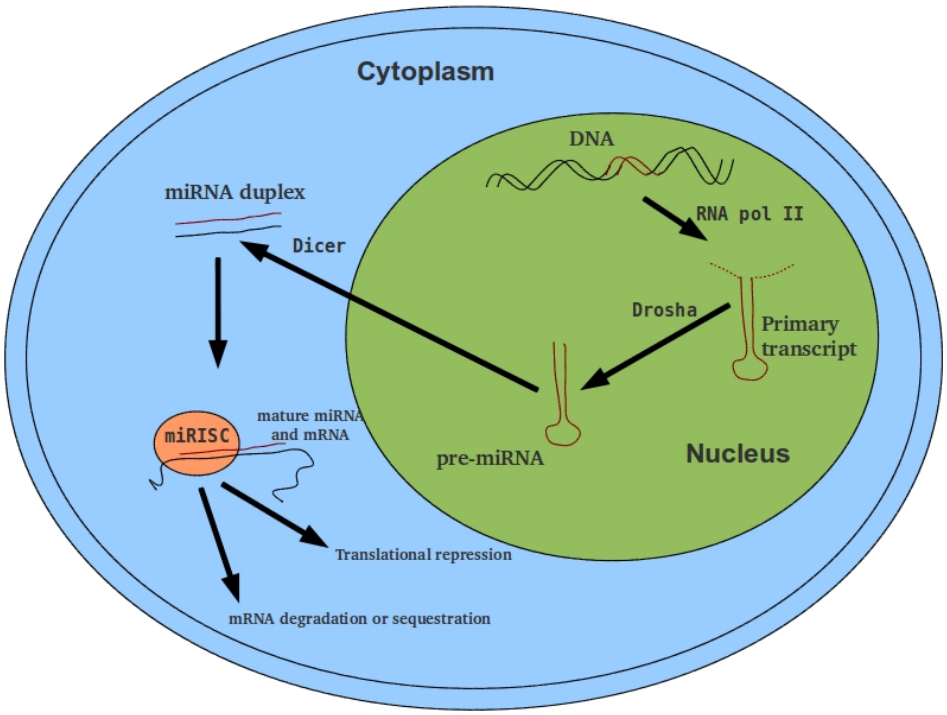


Fig. 1.7. miRNA biosynthesis. The primary transcript is processed by DROSHA into the harpin pre-miRNA. Then the fragment is exported to the cytoplasm and cleaved by another RNAse DICER into a duplex. The guide strand of the duplex binds with the miRNA induced silencing complex (miRISC) and then to the target mRNA. This results in transcriptional repression, degradation or sequestration of the target gene function.

1.4 miRNA and mRNA interaction

It has been discovered that the genomes of organisms contain hundreds of small non coding genes called microRNA (miRNA). miRNA are single-stranded RNA molecules of 21-23 nucleotides in length transcribed by RNA polymerase II. Primary transcripts are not translated into protein but are processed into hairpins by the RNAse enzyme Drosha, and exported into the cytoplasm, where they are cleaved by the central enzyme of the RNAi pathway, Dicer, to form single-stranded mature microRNAs [Amb04, Bar04]. Mature molecules have the function to downregulate actively gene expression via base-paring to the target transcripts [KHS09, WSGL07] (see Figure 1.7).

Plant microRNAs were recognized to function through near-perfect base pairing with their targets [LXKC02, TRBZ03] in the coding region and regulate the amount of transcribed protein via mRNA degradation [Wil08]. In animals, mature microRNAs bind to an imperfectly matched binding site in the mRNAs of target genes and regulate their post-transcriptional expression. In all known cases microRNAs repress expression of target genes, mostly by repressing translation while not a□ecting the mRNA concentration of the target, and less often by directly inducing a decrease in target mRNA concentrations[NMB+03].

1.4.1 miRNA mRNA association

The description of miRNA function depends on the target mRNAs and on the biological processes in which they are involved. The signal between these two level and that implement the miRNA cell function is poorly understood. A way to disentangle and understand correlative changes in miRNA expression and biological functions is to 1) follow reductions in the level

of a target protein (and by a potentially wrong assumption in animals reduction in target mRNA), 2) infer subsequent e□ects that may alter the levels of other mRNAs and 3) follow the ramification through the transcriptional profile to a whole biological process.

The first step is critical because in animals the regulated protein concentration is unrelated with the corresponding mRNA concentration. In the last several years, a number of di□erent miRNA profiling strategies have been used for following the miRNA expression changes in di□erent processes like di□erentiation, apoptosis, cell proliferation and a range of diseases. In this way it is possible to link directly miRNA expression with biological complex phenomena .

The most important strategies used are:

- *Mirna Q-RT-PCR Profiling;* a standard RT protocol is applied on a RNA sample and miRNA specific primers will probe for a specific miRNA through PCR amplification. This protocol is useful for a limited amount of RNA such as needle biopsies.
- *Array Platform*; probes for miRNA are spotted or synthesized on a chip. This has the disadvantages to require high amount of RNA.
- *Bead Based Method*; it requires flow-cytometry where beads with specific miRNA probes are attached and a mixture of two fluorescent dyes is used for coding uniquely the bead.

After these technical innovations, complex analysis in whole RNA permits the detection of microRNA signal correlating miRNA and mRNA profiles [LPP+07, WL09, FCC+09].

Chapter 2

Methods

2.1 High dimensional molecular data (HDMD)

High throughput technologies, like microarray, genome sequencing and proteomics, are generating large amounts of high dimensional data that have to be analyzed and understood. Properties of high dimensional data can affect the ability of statistical methods to extract meaningful information. For genomic and proteomic analyses, these properties reflect both statistical and mathematical features of high dimensional data spaces.

2.1.1 Some statistical aspects of HDMD

Usually many more data points than sampling units ($p \gg n$) HDMD typically has many more data points (D = number of variables or dimensions) than sampling units (n). Classical multivariate statistics were mostly based on assumption that D is fixed and sample size $n > D$ or even $n \to \infty$. Recent technological advances have generated much high dimensional biological data where $D \gg n$. Many multivariate statistical methods are impractical, fail, or are computationally not feasible when $D > n$. Multivariate statistics typically involves operations on a covariance matrix. When $D > n$, the covariance matrix is typically ill-conditioned, i.e., it is "singular" and does not have an inverse. Many statistical procedures, e.g., ML, can not proceed when singularity occurs. It is possible to use a generalized inverse in calculations but the ramifications of this are complex.

Meaningful replication is rare Replication is the repetition of the basic experiment. It is necessary to estimate the error variance and obtain a precise estimate of the effects of interest. Without replication, estimation of sample statistics is unstable, inaccurate and variation levels (technical, biological, environmental) within samples remains unknown. An example, evolutionary models depend on the understanding of the extent of intrinsic genetic variation in a sample; natural selection requires genetic variation to operate. Is the variation observed "biological?"

Initial data collection not hypothesis driven The search for critical variables to test hypothesis is complicated by presence of many irrelevant variables. The problem becomes how to separate critical variables from a vast pile of glut. This process introduces many complexities. Post hoc variable selection can be difficult. In designed experiments, it is assumed that the researcher is dealing with a few well-chosen variables. That is, scientific knowledge is being used to measure the right variables in advance.

Dimensionality of information unknown Gene expression data typically consist of a measurement of the (relative) expression level registered for a spot in correspondence to a particular hybridization. Formally, then, we have the matrix

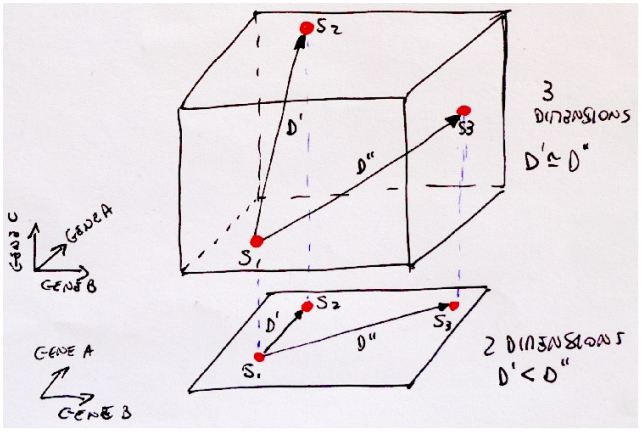


Fig. 2.1. Course of dimensionality. S_1, S_2 and S_3 are three experimental points projected in 2 (GeneA,GeneB) and 3 (GeneA,GeneB,GeneC) dimensional spaces. In 2D the distance $d(S_1, S_2) = D$ is smaller than the distance $d(S_1, S_3) = D$ whereas in 3D the two distances become equal.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

where the element x_{ij} is the measured expression for spot i in hybridization j. We will assume that there is one row per gene and one column per di□erent experimental condition, e.g., tumor type or cell cycle phase. The data matrix X has n rows and d columns. In classical statistical paradigm, *n* is in the order of thousands while *d* is in the order of tens. With HDMD, the conditions are usually reversed where $d \gg n$. Correlated variables means unknown dimensionality of actual information.

The "Course of Dimensionality" The notion of "Curse of Dimensionality" (COD) was coined by Richard Bellman (1961). It refers to the exponential increase in resources in computing a task of interest when extra dimensions are added to an associated mathematical space. For example, it arises in solving dynamic programming and optimal control problems when the dimension of the state vector is large. The amount of data needed to sustain a given spatial density increases exponentially with the dimensionality of the input space, i.e., the sparsity increases exponentially given a constant amount of data, with points tending to become equidistant from one another (Figure 2.1). In high dimensional spaces, data become extremely sparse and are far apart from each other. The curse of dimensionality a□ects any estimation problem with high dimensionality and it is a serious problem in many real-world applications like Microarray data (3,000-30,000 genes).

2.1.2 Dimensionality reduction

Taking high dimensional interrelated data, the goal is to reduce dimensionality (d) of a data set (sample) by finding a new set of variables (k) smaller than the original set of variables ($k < d$), that nonetheless retains most of the sample's information; we can:

- Reduce the dimensionality of problems
- Transform interdependent coordinates into significant and independent ones

2.1.3 Principal component analysis (PCA)

This method is a classical application of the Single Value Decomposition (SVD). The goal is the compression of redundant matrices and the identification of important properties inside a measurements matrix A ($m \times n$). If we use the covariance matrix (assuming that the matrix A has mean value equal to 0)

$$COV(A) = \frac{1}{n-1}(a_1 a_1^T + ... + a_n a_n^T) = \frac{1}{n-1} AA^T$$

the o-diagonal elements show correlations between samples.

The AA^T matrix is the correlation between rows (usually samples in microarray experiments) and the $A^T A$ matrix is the correlation matrix between columns (usually genes). As it is known from linear algebra the SVD returns the fundamental factorization $A = U\Sigma V^T$ where U and V^T are two orthogonal matrices and Σ is a diagonal matrix that contains singular values. Is easy to find that $AA^T = U\Lambda U^T$ and $A^T A = V\Lambda V^T$, where $\Lambda = \Sigma\Sigma^T$. So U and V are two orthogonal basis for AA^T and $A^T A$. The rank of A gives the number of the r first independent eigenvectors in U and the first r eigenvalues of U are the best basis in the sample space $\mathbb{R}^m$.

Usually the matrix $AA^T(A^T A)$ has full rank but comparing the singular values $\sigma_1, \sigma_2, ..., \sigma_n$ we could determine how many combinations of samples are revealed by the analysis. Usually, the rank r is empirically selected as the index of the score in Σ that had a value below a threshold (threshold= 1; Keiser Criterion). For example if $\sigma_4 < thr$ then $rank(A) = r = 3$.

2.1.4 Factor analysis (FA)

Factor analysis was initially developed by Charles Spearman in 1904 and the term was first introduced by Thurstone in 1931. Because FA and PCA share some operational procedures, they are often confused in the literature. PCA describes the patterns of observed variation in a sample whereas FA relates the covariance patterns to shared co-variability related to an underlying latent structure.

2.1.4.1 Correlation matrix The main difference between FA and PCA is that conversely to PCA, FA has not an unique solution. This is due to the use of a slightly different version of the correlation matrix used in PCA and FA. In PCA the diagonal of the matrix (a_{ii}) is filled with '*1s*', (this feature of the classical correlation matrix drives to an unique solution) whereas in FA, the diagonal of the correlation matrix has the '*1s*' replaced with estimates of communality that measure the variation of a variable in common with all the others together. This estimating process makes the solution of FA not unique (and different from PCA). Most of the algorithms employed for the communality measure are based on the *squared multiple correlation coefficient* (SMCC) of the variable in the diagonal with all the others.

2.1.4.2 Factors rotation The principal axis theorem on which FA and PCA are based returns an orthogonal factor matrix. With this factor matrix we have reduced the dimensionality of the problem but the factors may be not easy to interpret. As showed in table 2.1 two factor matrices are returned (usually) with FA, one unrotated and one rotated. The unrotated matrix comes directly from the principal axis theorem (or spectral theorem) that states that every symmetric matrix A has the factorization $A = Q\Lambda Q^T$ where Λ are real eigenvalues and Q are orthonormal eigenvectors. The unrotated output makes FA and PCA similar (and confusing) because results are obtained using analogous mathematical tools (principal axis theorem).

Rotation facilitates the interpretation of factors and makes FA output more understandable. Rotations are grouped in two classes: orthogonal and oblique (see Figure 2.2) .

Orthogonal Orthogonal rotation assumes that the factors are independent and the rotation process maintains the reference axes of the factors at 90 degrees.

	Unrotated factors			Rotated factors		
variables	1	2	3	1	2	3
variable 1	*0.96*	-0.03	-0.4	*0.65*	*0.62*	-0.01
variable 2	*0.93*	0.12	-0.05	0.32	0.12	*0.71*
variable 3	0.39	*0.84*	0.26	0.39	*0.77*	-0.16
variable 4	*0.95*	-0.21	-0.07	*0.75*	-0.34	0.48
% common variance	51.9	38.1	10	40.6	35.3	24.1

Table 2.1. Rotated and unrotated factor matrix. Columns define factors, rows variables. The number of factors is the number of independent patterns of relationship in the data. The first factor in the unrotated factor matrix contains the largest pattern of relationships (*italics*). The second one delineates the next largest pattern and so on. The total amount of common variance described by each pattern decreases successively. In the unrotated matrix factors are unrelated with each other. In the rotated matrix the patterns of relationship are distributed over all the factors. Factor 1 and 2 are related with variable 1.

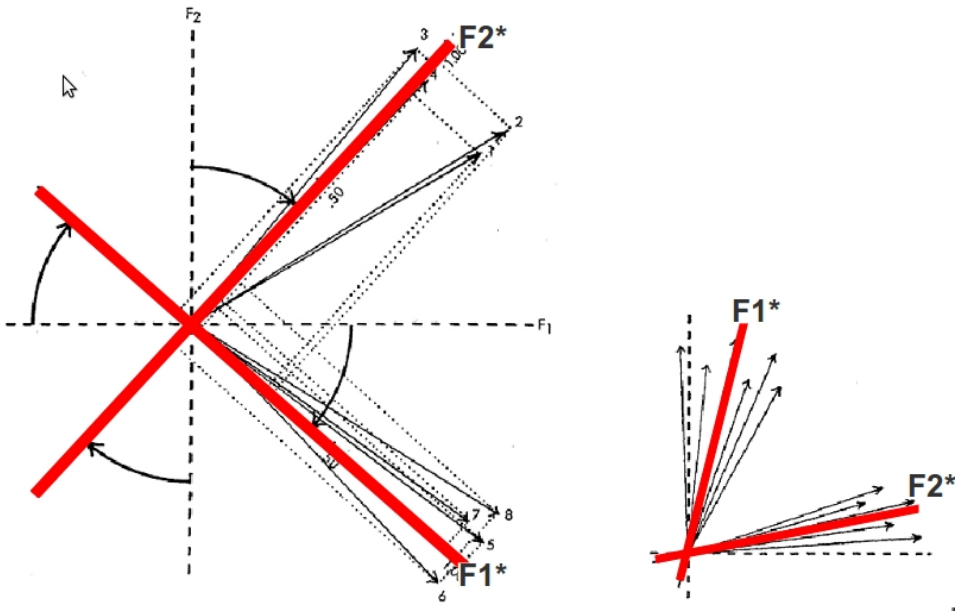


Fig. 2.2. Two factors structure rotation; red lines indicate the new position of the rotated factors F1* and F2*. Left: orthogonal structure rotation. Curve arrows show direction and degree of rotation, straight arrows show variables. Right: oblique rotation of factors that fit two clusters of variables (straight arrows).

Oblique Oblique rotation does not require the reference axes to be at 90 degrees. This rotation may be more realistic representing the clustering of the variables more accurately.

The rationale is that rotations simplifies the factor structure and makes interpretation easier. Thurstone [THU46, THU47] suggested five criteria to identify a simple structure. A matrix of loadings (where rows correspond to the original variables and columns to the factors; see table 2.1) is simple if:

1. Each row contains at least one zero (or a small value).
2. For each column, there are at least as many zeros as columns.
3. For any pair of factors, there are some variables with zero loadings on one factor and large loadings on the other factor.
4. For any pair of factors, there is a sizable proportion of zero loadings.
5. For any pair of factors, there is only a small number of large loadings.

All the various rotation methods have, as a guiding principle, the simple structure concepts and the results, after rotation, should have become simple in their appearance

2.2 Next generation sequencing

The first two complete genome sequences of living systems became available in 1995 and were the genomes of two bacteria: *Mycoplasma genitalium* [FGW+95] and *Haemophilus influenzae* [FAW+95]). Since that year more than 800 genomes were sequenced providing a huge amount of material for comparative analysis.

As stated by von Hejne [Hei87]

> *...molecular biology is all about sequences. First, it tries to reduce complex biochemical phenomena to interactions between defined sequences...*

the assumption of bioinformatics is to consider di□erent biological problems as problems defined on sequences (or strings). In the genomic era, after more than 40 years since the first global alignment algorithm was published [NW70] and in which new sequencing technologies are able to produce Giga bases per day the central challenge in computational biology is still sequence alignment.

2.2.1 Sequence comparison

The interest in developing computer algorithms to analyze DNA sequences produced with new sequencing technologies increased in the 1970s.

In 1970 a method based on matrices was described to compare two DNA sequences [GM70]. At the same time Needelman and Wunsch [NW70] started a detailed examination on the sequence alignment using dynamic programming. This idea was the seed for a plethora of methods that where developed in the in subsequent years.

2.2.1.1 Distance between two sequences There are several ways to assign a distance between sequences. The simplest one is to use an edit distance. It is a distance in which each edit operation on single characters that transforms one string to the other string is defined with a score. The edit operations are:

- *Match (M)*; corresponds to identity or to non operation.
- *Substitution (S)*; is a replace operation.
- *Insertion (I)*; is the insertion of a character in the first string.

- *Deletion (D)*; is the deletion of a character in the first string.

An insertion in the first sequence is a deletion in the second and the opposite. For example if $s1$ = $ACTTACTCAG$ and $s2$ = $actgtgctat$ the sequence $s1$ can be edited to become $s2$ as follows

```
ACT TACTCAG
MMMIMSMMDMS
actgtgct at
```

So we can define the edit distance between two sequences $s1, s2$ as the minimum number of edit operations needed to transform $s1$ in $s2$.

2.2.1.2 Dynamic programming approach The edit distance problem could be solved via dynamic programming. The main idea is that if we have two sequences $s1, s2$ the minimum number of edit operation needed to transform the first i letters of $s1$ and the first j letters of $s2$ is called $D(i, j)$. So if the two sequences have length m and n respectively the edit distance between $s1$ and $s2$ is exactly $D(m, n)$.

The dynamic programming approach computes $D(m, n)$ computing all the $D(i, j)$ for each the combinations of i and j starting from 0 to m and n.

2.2.2 Local and global alignments

When we compare two sequences the two strings may be or not similar in their entirety. Global alignments expect that the whole sequences are similar and compute the edit distance on the complete length of the sequences. On the contray local alignments find and compute edit distance on regions that exhibit similarity.

2.2.3 The Giga base era

All new sequencing technologies in production, including 454, Illumina, SOLiD and Helicos, are able to produce data in the order of giga base-pairs (Gbp) per machine/day. Researchers have quickly realized that even the best tools for aligning capillary reads are not e□cient enough given the unprecedented amount of data and many new alignment tools have been developed in the last few years. Most of the fast alignment algorithms construct auxiliary data structures, called indexes, for the read sequences or the reference sequence, or sometimes both. Depending on the property of the index, alignment algorithms can be largely grouped into two categories: algorithms based on hash tables, algorithms based on su□x trees.

2.2.3.1 Algorithms based on hash tables BLAST was the first aligner that used the idea to construct hash tables to speed up the alignment. All hash table based algorithms essentially follow a seed-and-extend paradigm. They keep the position of each k-mer sub-sequence of the query in a hash table with the k-mer sequence being the key, and scan the database sequences for k-mer exact matches, called seeds, by looking up the hash table. For example, given the reference sequence `ACTCACTGACTT` and the query `actg`, an algorithm with k = 3 builds on the query a hash table with two components `act`=000111 and `ctg`=011110. In table 2.2 it is showed how the scan works.

Then for each match the algorithm extends and joins the seeds first without gaps and then refines them by a Smith–Waterman alignment.

spaced seed Ma et al. [MTL02] discovered that seeding with non-consecutive matches improves sensitivity. For example, a template '111010010100110111' requiring 11 matches at the '1' positions is 55% more sensitive than BLAST's default template '11111111111' for two sequences of 70% similarity. The template with internal mismatches is called spaced seed. Most of the

k-mer	Hash code	Match
ACT	000111	*
CTC	011101	
TCA	110100	
CAC	010001	
ACT	000111	*
CTG	011110	*
TGA	111000	
GAC	100001	
ACT	000111	*
CTT	011111	

Table 2.2. Hash table for the reference sequence ACTCACTGACTT. In this example 4 matches are found (*). The two central matches combined together give a perfect match for the query on the reference. The hash code is an integer composed by six bits, two bits for each base using this schema: A=00; C=01;G=10;T=11.

Program	Data Structure	SOLiD
BFAST	HASH	Yes
BLAT	HASH	No
MAQ	HASH	Yes
MEGABLAST	HASH	No
MOSAIK	HASH	Yes
BOWTIE	Su□x	Yes
BWA	Su□x	Yes

Table 2.3. Short reads aligners.

tools that use spaced seeds allow k-mismatches. For instance, MAQ [LRD08], builds six hash tables to index the reads and scans the reference sequence against the hash tables to find the hits. It ensures that the first 28-bp of the reads with two mismatches or fewer will be hit.

With this kind of methods it is possible to align a query on a L-long target sequence in $O(kL)$ time, where k is the number of mismatches, independently of the length of the query.

2.2.3.2 Algorithms based on suffix trees A su□x tree is a data structure that permits to solve problems related to pattern matching in linear time. The classical application is to find the sub-string s in a text T in a time that is $O(length(s))$ regardless of the length of T.

A simple su□x tree is shown in Figure 2.3 where a naive procedure is used only to give an intuition of the data structure. Methods that implement su□x trees and their modifications could search the N-long query in $O(N)$ after that the su□x tree, for the L-long target, is built in $O(L)$. However a classical su□x tree takes $O(L^2)$ space making the tree construction impractical. Improvements on data structure compression achieve linear spaces and linear time search.

The popular aligners for short reads are presented in Table 2.3

2.2.4 SOLiD reads

As it was described in 1.3.1.3 the SOLiD technology reports two adjacent bases at the same time and a single color sequencing error makes the following sequence wrong. For this reason algorithms that naively decode a color read will fail. The optimal approach is to revert the target sequence in color-code, align reads and target in the color space and then convert the result of the alignment into base code. Most of the tools use this strategy to align SOLiD reads to target (see Table 2.3).

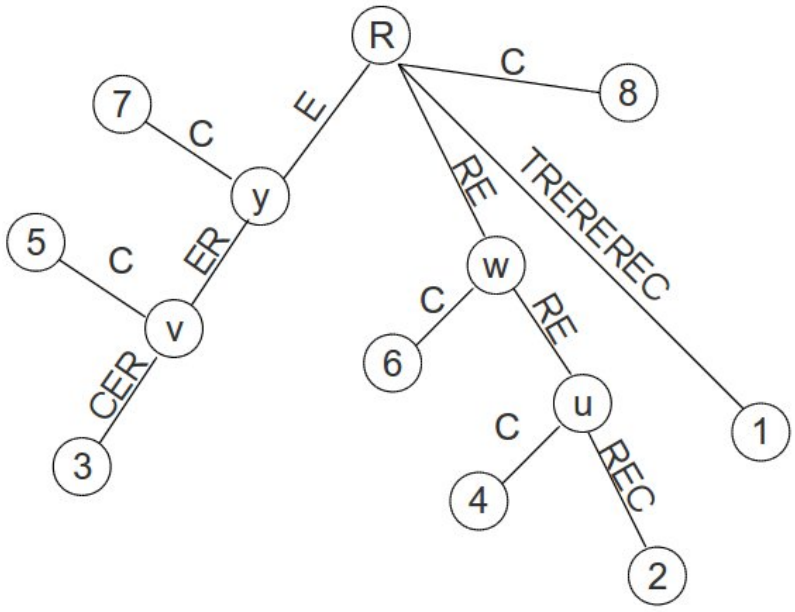


Fig. 2.3. Suffix tree for string s ="TREREREC". First the method enters a single edge (node 1) for the complete suffix $S(1, m)$ where m is the length of s; then enters suffix $S(i, m)$ into the tree for i that goes from 1 to m. Edges are labeled with the string that is found in the longest path from the root (R) whose previous label matches a prefix of $S(i+1, m)$. When no further matches are possible and the suffix $S(i+1, m)$is not finished, the algorithm breaks the previous edge and inserts a new node (nodes u, v, w and y) just before the first character on the edge that mismatched.

Part II

Projects

Chapter 3

In silico SNP detection, probes design and Illumina's Porcine Chip annotation

3.1 Description

The lack of a pig genome assembly, the need of a procedure to map an obesity-related human set of genes (ORH) on the porcine coding-space and the request to expand the ORH into a new ORH-related set (OR^2H) hampered the pipelines already developed for SNPs discovering.

Any tool used in SNPs discovery contains at least two main components: 1) one methods to align short sequences of different individuals to a reference genome using alignment algorithms [MM88] and 2) a system that scan the alignments in order to find SNP candidates (PolyPhred [NTT97], Maq [http://maq.sourceforge.net/index.shtml] and others). Many complex pipeline was developed to detect SNPs in pigs, using EST [PSH+07, TVV+06, TLV+08] or using a genome-wide approach [KKK+09]. All these systems are unusable for our aims since they can not exploit, in the same time, information that arise from sequences that come from two different species. Moreover their complexity complicates the re-parameterization of the tools making existing pipelines useless for our specific goals.

To overcome the constrains imposed using "general purpose" pipelines we addressed the request formulated in the project:

1. designing a comparative and systemic approach that uses the human interactome to select and recruit genes that are linked with genes in ORH;
2. extracting and implementing a medium density chip in which a set of genes potentially involved in fat traits regulation are included;
3. assigning a possible significance to most of the SNPs included in the new Illumina's PorcineSNP60 BeadChip array that will be used to address the study from a genome wide point of view.

The first two steps were performed at the beginning of the PhD when the complete assembly of the pig genome was not disposable while, after the release of the pig genome in 2010, the third task was executed.

3.2 Platform implementation

The detection of the SNPs was obtained analyzing public sequences achieved using capillary DNA sequencing machines, such as the ABI PRISM 3100. These machines generate vast amounts of raw sequence data with little user intervention. So we developed a framework to facilitate automated DNA sequence analysis connecting most of the standard software used to analyze such kind of data. The implementation is based on a standard 3 independent layers architecture: 1) a data layer; 2) a logical layer and 3) a presentation layer (Figure 3.1).

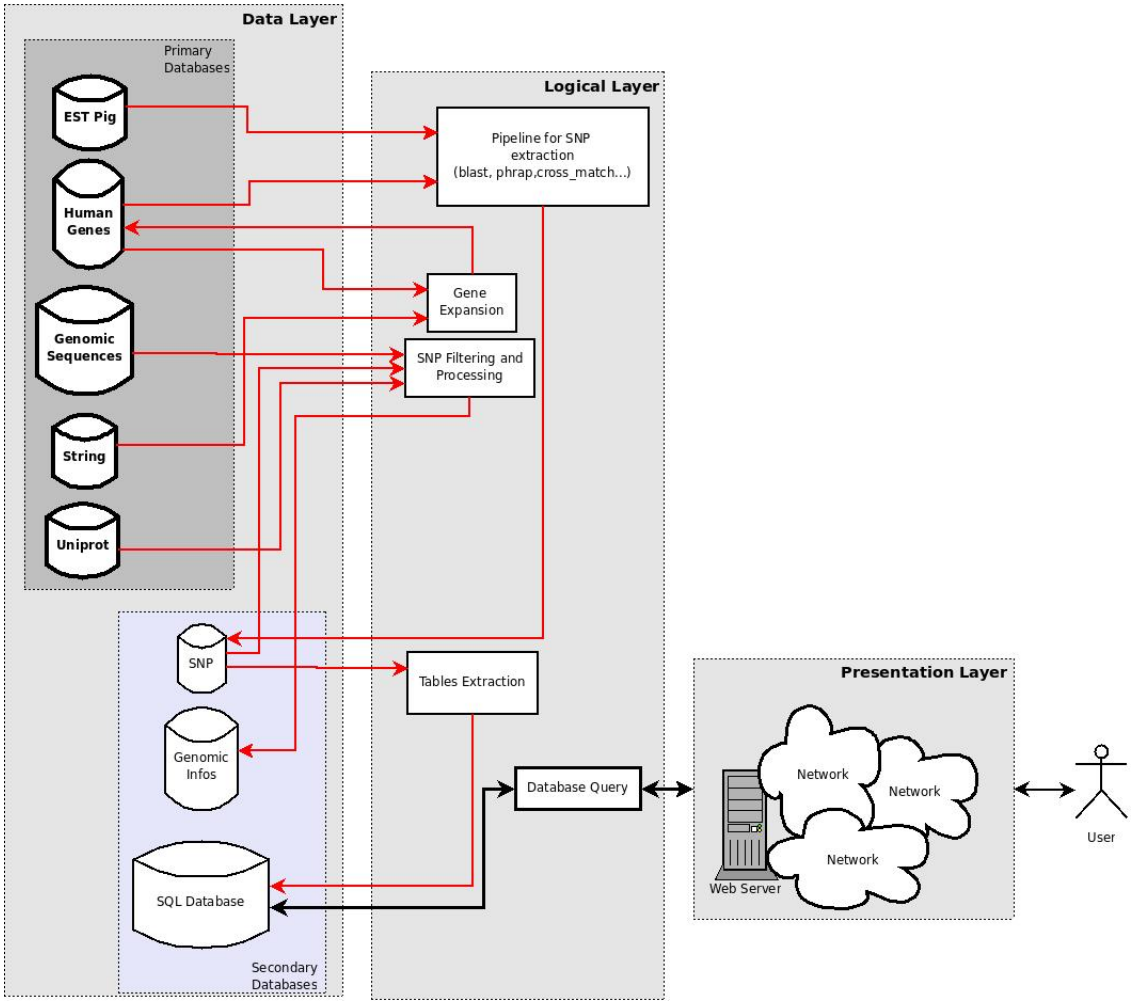


Fig. 3.1. The platform for SNP detection in animal models.

3.2.1 Data layer

The data layer contains the datasets used for SNP detection/annotation and gene expansion and is responsible to store, maintain and provide information e□ciently. Data-sets are saved for convenience in a mySQL database server. The schema of the tables was kept identical to the schema provided by the maintainers of the database from which tables are from. Custom tables are designed in order to achieve a normal form.

Data-sets Expressed sequences (2,224,772) and EST trace records (1,035,200), consisting in over 2 billion of bases, were retrieved from ncbi using QUERY_TRACEDB (http://www.ncbi.nlm.nih.gov/Traces/). The trace sequences derive from two cDNA libraries, the Sino-Danish Joint Venture Partnership (http://www.piggenome.dk/)and the US Meat Animal Research Center (http://www.ars.usda.gov/). All the sequences were retrieved in fasta format. Human genes (more than 5,000 cDNA sequences) were obtained from ENSEMBL (http://www.ensembl.org/) via Biomart (http://www.ensembl.org/biomart/martview) using symbols curated by the HUGO Gene Nomenclature Committee at the European Bioinformatics Institute (HGNC) (http://www.genenames.org) to query the database. 57,444 pig partial genome fragments are downloaded from NCBI whereas assembled chromosomes was obtained from ftp.ncbi.nih.gov/genomes/Sus_scrofa. The STRING proteins network (version 8.5) was downloaded from EMLB after license request (http://string.embl.de). susScrofa genome version 9 (April 2010) and the new version 10 (September 2010) were the downloaded from http://hgdownload.cse.ucsc.edu/downloads.html.

3.2.2 Logical layer

It is responsible in the data processing and contains the programs and the scripts that process information provided by the data layer. It consists of three main components: 1) a set of scripts for gene expansion; 2) a pipeline for SNP detection and 3) a subsystem for SNP filtering and processing. The three components are embedded in a pipeline that coordinates the processes and the information flow.

3.2.2.1 Expander A local version of the STRING proteins network was moved into our mySQL database. HGNC gene name convention is used to query the database. The fasta format of each interacting protein is fetched from a local selection of the ENSEMBL 53 database.

STRING uses eighth di□erent scoring systems that come from:

- literature (Textmining score);
- experimental data (Neighborhood,Gene Fusion, Cooccurence, Coexpression and Experiments score);
- biochemical networks (Databases);
- a resume of the other seven scores (general score).

The general score goes from 0 (no correlation) to 1 (certain correlation). Two genes were considered as associated when the general association score between them is above 0.8[1].

[1]The selection of this threshold was obtained after the following simple procedure:

1. a random sample of 30 genes was selected;
2. the random set was expanded using different thresholds (from $t_0 = 1.0$ to $t_9 = 0.2$, step 0.1);
3. the number of the returned "interactors" was plotted;
4. starting from the more stringent search $t_0 = 1.0$, the threshold was selected manually searching approximately the point p_t where the number of returned genes g_{t_n} differ a lot from the number $g_{t_{n-1}}$.
5. The threshold t_{n-1} was kept.

The resulting threshold is really stringent compared with the default confidence threshold (0.4) used by the STRING web service.

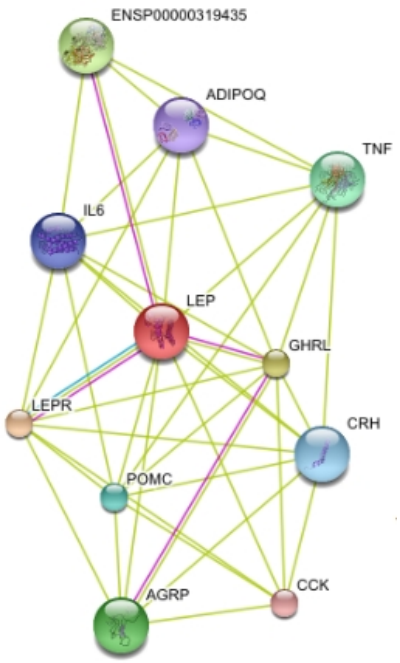


Fig. 3.2. Gene-subnet around LEP (leptin) in STRING. Proteins are represented by nodes and scores by edges. Marked edges represent relations where scores are above a threshold. In this picture all the proteins are directly connected with LEP with at least one score; GHRL, LEPR and ENSP00000319435 are linked with LEP with two or more scores. In our system I used only the general score that resume the other 7 scores.

Briefly, a query procedure implemented in PYTHON takes a list of genes and return a set of genes that shares a general association score above 0.8 (Figure 3.2). In our case the input gene list is the ORH set. In this set each gene is labeled using the HGNC nomenclature whereas in STRING each node in the interaction map is labeled with a unique ID that represent the ensembl Proteins id. To allow the use of the standard human gene nomenclature a dictionary provide a reversible conversion between HGNC name and ensembl Proteins id.

ORH contain a non-random genes sample that belong to the obesity pathways. After the expansion procedure connections between genes in the ORH are highlighted and a new set, the OR^2H, has been obtained. As shown in Table (3.1) more than a half of the genes in the ORH are covered by the OR^2H. As a matter of fact, 163 genes (71.5%) in the ORH set were retrieved via expansion procedure. The probability that an expansion procedure, with a random set of genes, retrive the same fraction of original genes approach to zero.

3.2.2.2 Contigs reconstruction The pipeline was implemented on a Linux system based on a 2x4 cores architecture. It was developed using a mix of shell scripts and Perl modules. The scripts have been used to connect and coordinate several packages freely available for academic use.

We divided the identification of raw SNPs in the obesity related gene sets (ORH and OR^2H) in a 2-step process:

1. BLASTN [AMS+97] has been used to align the human genes to the EST dataset and obtaining, for each human gene, a cluster of orthologues porcine ESTs.

2. REPEATMASKER [SHG] was used to remove portions of low complexity DNA from the clusters and PHRAP/CROSSMATCH [GAG98] have been used to assembly EST clusters (`http://www.phrap.org`). Alignments and contigs are returned in .ACE format (Figure 3.3) where each sequences cluster (putative porcine gene) is saved in single file. ACE files

```
QA 1 356 1 356 DS rmed08c_p23.y1
RT{ gi|18283.. matchElsewhereHighQual phrap 36 356 090304:170210 }
RT{ gi|18281.. matchElsewhereHighQual phrap 1 292 090304:170210 }
CO Contig14 1102 31 193 U
aaATCCCAAGGGCAAGAGCTGCACACTTACTGCCTGTGCTCCTTCCCAGC
TGGGGCTGAGGGCCACG*GGAGGGGCAACTCTGGAAATAAAGTGGAAGGA
GGGCTGGGCTGAGCCGGGAAGTGTTGAAAATCTGGGAGGCTGAGAGTGGA
...
ATCCCacacccccctgaccCCCacaggACCATGAATGaggaagcaagggga
acggctggaaatctgggctgggttggacaaaggacaactgagggttcttg
BQ 0 0 50 50 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60
60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60
60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60
60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 .. ..
...
```

Fig. 3.3. Contigs in ACE format. This text format is space consuming and is not easy to manage. A parser has been created to extract and provide SNPs information to the next steps.

```
GE ABCG5_44.screen.rel
NA Contig1 1 1
NA Contig2 2 2
NA Contig3 4 4
//
GE ABCG5_513.screen.rel
NA Contig1 1 1
NA Contig2 2 2
NA Contig3 4 4
//
GE ACE_111.screen.rel
NA Contig1 1 1
NA Contig2 1 1
...
NA Contig9 7 7
NA Contig10 8 8
SQ gCCACATTCCTTCAAGTGTGCCTTACATCAGGTACTTTGTCAGCTTCATCATCCAGTTCC...
PR ATFLQVCLTSGTLSASSSSSSSMRRCARRXGHKXPPAQV*QLPVQGSREAPGGRHEAGPQ...
OR gnl|ti|1377353023 gi|182503309|gb|EW285412.2|EW285412 gi|153...
SN 123 GGGTTTT
PS _____________________________________________________V
FR CGCTGTGCCAGGCGGCG*GGCCACAAA*GGCCCCCTGCACAAGTGTGACAGCTACCAGTC...
NA Contig11 37 37
//
```

Fig. 3.4. Flat text format returned by the .ACE parser. Each row start with a field with two letters. *GE* stands for GENE and flag the beginning of a new record for the gene; *//* is the end of the record. *NA*, indicate a new contig. If a contig contains one or more variations the NA field is followed by: *SQ*, the contig sequence; PR (optional), the sequence of the translated protein; *OR*, the ID list of the EST that originate the contig; *SN*, the column vector in the alignment that contains the variation; *PS*, the position of the SNP on the fragment *FR* (101 bp length) that contains in the middle the variation. In this example 3 genes are listed. The first 2 isoforms of *ABCG5* contain no variations. In the gene *ACE* a variation is found in Contig10 (composed by 8 EST). From the SN field we see that in position 123 (Contig10 coordinate) the variation contains 3 G and 4 T.

are parsed by a home made script that extract raw SNPs from the contigs and return flat files (Figure 3.4) that contain information in a format easily manageable by humans or parsers.

3.2.2.3 Filtering and processing module Our procedure is a□ected by artifacts deriving from a well defined source of noise. This noise is introduced by the clustering procedure owing to the inter-species paralogs/pseudogenes. Such artifacts are composed by contigs containing regions with high density of SNPs. We implemented a fast strategy assuming that the frequency of SNPs in coding regions is one mutation every 300 bases [RGA03] and that the SNP distribution on DNA may be characterized, at first approximation, by a Poisson process [KN01].

For each pair of SNPs we compute the probability $p(n-(n+l)) < 2$ according to the relation

$$p(n-(n+l)=k) = \frac{e^{-\lambda l}(\lambda l)^k}{k!}$$

where $n-(l+1)$ is the number of SNP in a sequence of length l, λ is the mutation intensity parameter (1 mutation each 300 bases) and $k(=2)$ is the number of observed events. So the probability that we observe two or more mutations , under the hypothesis of a poissonian process in a DNA stretch of length l, is

$$p(n-(n+l)\geq 2) = 1-(e^{-\lambda l}(1+\lambda l))$$

If the returned probability p is under 0.05 we reject the hypothesis that two consecutive SNPs follow the selected piossonian model and the two mutations are labeled as "bad" (as "good" on converse).

For each SNP labeled as "good" we attempt to annotate the following information: a) the putative gene name, b) the contigs number and the positions in which the SNP is found, c) the variation introduced, d) the minimum allele frequencies (MAF), e) the minimum allele count (MAC), f) the deep of the alignment, g) the sequences sources id, h) a genomic sequence centered on the SNP for genotyping and i) the sources of the pig genome fragments used to extract genomic sequences.

Furthermore, on each polymorphic location a score was evaluated to weight the "importance" of a SNP on the basis of the number of EST that contain the mutated base (MAC number). We calculated the probability P_{SNP}, under a simple binomial model, that at a specific location the transition from base X to base Y ($X \rightarrow Y$) occur in at least k sequences in an alignment of n EST, where k is the minimum allele count (MAC) of the polymorphic site. Under these assumption the probability that a SNP appear by chance is

$$P_{SNP} = 1-\sum_{i=0}^{k-1}\binom{n}{k}P_{X\rightarrow Y}^{i}(1-P_{X\rightarrow Y})^{n-i}$$

where $P_{X\rightarrow Y} = 0.0033$ is the probability that a base X is mutated. This score has been used to sort variations on the basis of the number of animals that bring the variant and to help the choice of the SNPs that will be used to design the custom chip.

3.2.2.4 Probes and chip design The top of the iceberg under my responsibility was the release of portions of genes that contain SNPs helpful to design the GoldenGate Illumina chip. A mix of methodological and technological constrains could reduced the number of the SNPs potentially useful in the chip design:

- Use of human genes to extract porcine EST. The base of our argumentation is that humans and pigs are species close enough to assign a porcine homologous to each human gene.
- Paralogs noise. Due to the previous point the number of paralog ESTs processed in the pipeline is higher than the number of paralog EST expected using genes that arise from the same organism.
- Low coverage. The mean coverage is about $\frac{3.3\,10^6\cdot 500}{1.5\,10^8} = 11$, where $3.3\,10^6$ is the number of ESTs, 500 is the mean length of each EST and $1.5\,10^8$ is the number of bases that belong to the trascriptionaly active part of the pig genome.
- Lack of an assembled genome. Only recently two versions of the pig genome were released. No information about intergenic/UTR/introns/pseudogenes regions was provided to the researcher before 2010.
- Length constrains to design Illumina probes (200 bases). Genomic DNA samples obtained from Italian Large White pigs has been used for genotyping experiments thus information about introns must be provided for probes design. This request was the strongest limitation because the contigs extracted were build on ESTs that contain only transcribed DNA.

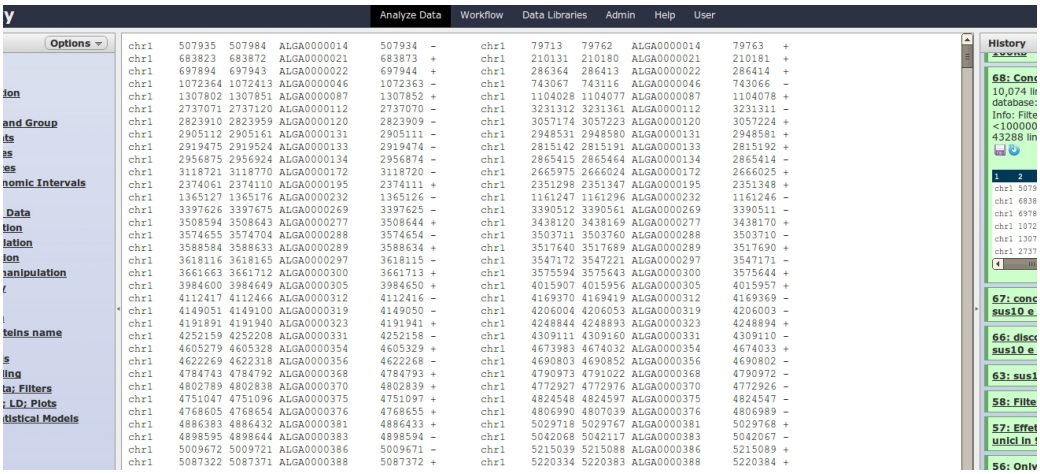


Fig. 3.5. Galaxy instance that compare the alignment obtained using two different versions of the pig genome of the probes present in the 60K SNP chip.

We exceeded, partially, the limitations in the last point obtaining information on pigs exons/introns gene structure querying chromosomes and BAC clones with our contigs via BLASTN. A portion of 400 bases is returned for each alignment longer than 20 bases[2] and with identity score above 96%. Because the portion that contains the SNPs could be located nearest an exon/intron splice-site a 20 bp long selection is required to avoid a lack in sensitivity.

3.2.3 Presentation layer

The results provided by the logical layer was displayed on a local instance of Galaxy [GNT10]. This platform is based on a complex server architecture where di□erent mechanisms are employed to compose a rich application server in which users are free to analyze/query/share/modify biological data (see figure 3.5). We developed several wrappers in order to control via Galaxy a wide spectra of tools.

3.2.4 String expansion

In our test case the starting candidate gene set (ORH) include 249 human genes (selected from literature and from the Gene Obesity Map, `http://obesitygene.pbrc.edu`) that has been expanded to include some other sequences from the human interactome (OR^{2}H; Table 3.1). From 249 genes 229 (91.6%) are found have a corresponding protein in String. Using the proposed threshold we expanded the obesity gene set to 2,431 genes where 229 are still known and 2117 are completely new. Figure 3.6 shows the view (presentation layer) of the implemented tool to control the expansion procedure. The tool is embedded in a local instance of the Galaxy Framework.

3.2.5 SNPs search

Results are kept separate for genes in the ORH and OR^{2}H sets (see Table 3.2). For both sets searching procedure has been performed in contigs derived from 3,259,972 pig EST. The distributions on the number of ESTs in contigs for both the dataset are shown in Figure 3.7.

[2]Roughly speaking I adopted this threshold (20 bases) because it gives, under an uniform model, 1 false positive alignment each $\frac{4^{20}}{500 \cdot 10^6} = \frac{2^{40}}{500 \cdot 10^6} \approx \frac{1000 \cdot 10^9}{500 \cdot 10^6} = 2 \cdot 10^3$ alignments, where $500 \cdot 10^6$ is the number of bases that compose the queried database.

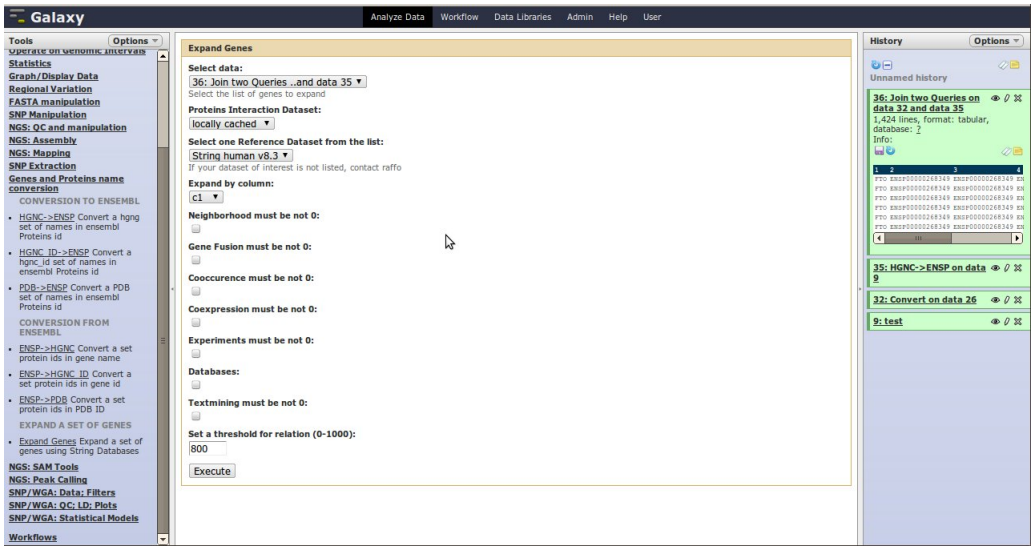


Fig. 3.6. The page is divided in three parts. The middle one is where the user can choose databases, genes list and parameters for the expansion procedure.

Input Data		*Number*
ORH[a]		
	Total	249
	Unique[b]	86
OR^{2}H[c]		
	Total	2,431
	Unique	2,202
Common		163

Table 3.1. Expanded Gene Set statistics. [a]From literature and database search; [b]for 20 genes in the ORH set we did not found correspondence in STRING; [c]using a threshold > 0.8 for the global score.

Input Data	*Output Data*	*Number*
Pig EST sequences[a]		3,259,972
ORH		249
	Contigs [f]	3,191
	Raw SNPs[b]	7,221
	Filtered SNPs[c]	1,583
	Annotated SNPs[d]	733
OR^2H[e]		2,202
	Contigs	29,018
	Raw SNPs	60,251
	Filtered SNPs	12,159
	Annotated SNPs	6,393
	Genes with probes	690

Table 3.2. SNP statistics of the test case. [a]from NCBI; [b]after the pipeline stages; [c]using the filtering procedure based on the Poisson process with $\lambda = 0.0033$; [d]at least a genomic fragment of 400 bases is associated with the SNP; [e]after a search in the human interactome of String; [f]the number refers to contigs composed by al least 4 EST.

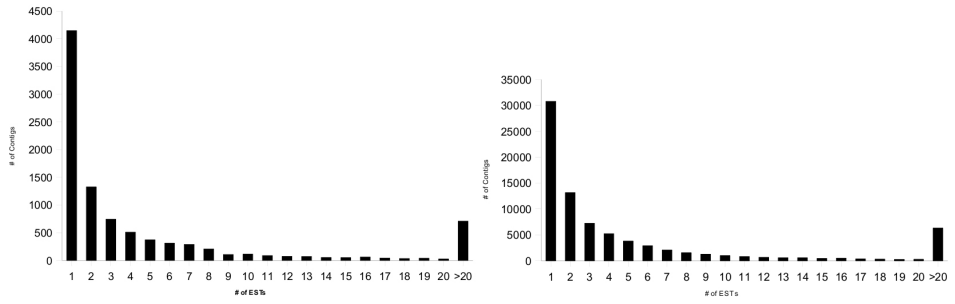


Fig. 3.7. The distribution of the number of ESTs per contig in A) the ORH and in B) OR^2H genes.

The first stage of the pipeline identified 7,221 bi-allelic raw SNPs from 3,191 contigs found in the ORH set . The contig assembly process on the OR^2H set gives 60,251 raw SNPs that, after the filtering procedure have been reduced to 12,159 (20.2%) mutation. The polymorphism rate obtained is $\approx$ 1 mutation each 800 base pairs and every gene sequence contains on average 3.6 SNPs. This result approximate the right bound (1/1000) in the expected SNP density interval of a mammalian genome. The small mutation rate could be explained with some considerations.

First, we search SNPs in coding regions were the selective pressure is high and the mutation rate is lower than in non coding region [ZFHEB03] and second, the low coverage of the EST on the porcine genes ($\approx 11\times$ and 0.32 for coverage and MAF, on average),reduce the sensitivity of the method that return variations that appear in a small population's fraction. As shown in Figure 3.7, 95% of the polymorphisms have less than 9 ESTs that bring the alternative allele (MAC) and about the 56% of them (removing singletons) have MAC that is ≤ 2 . This value is useful to estimate the signal to noise ratio: higher is the number of EST aligned with the alternative mutation and lower is the probability that the mutation is due to random errors in the EST sequencing procedure. This observation driven the decision to use the binomial model showed in section 3.2.2.3 to score each filtered SNP with the probability that the variation is accidental.

Here we want to start a short discussion on the issue of the SNPs isotypic multiplicity. We are using a set of 2,431 human genes mapped on a space of 5,290 sequences; this implies that each gene is represented, on average, by $\approx$ 2.2 isotypic sequences. Identical SNPs found

GENE	*PP*	*HS*	*Common*	*Experimental*
ADRB2	[b]0 [c](0)	0 (0)	0	1
AMPD1	2 (1)	0 (0)	0	1
BDNF[a]	1 (1)	0 (0)	1	1
CIDEC[a]	1 (1)	1 (1)	0	1
CSTB	0 (0)	1 (1)	0	1
CTSL1	14 (1)	13 (1)	1	1
CTSS[a]	6 (1)	0 (0)	0	1
FTO	2 (0)	0 (0)	0	1
LEF1	3 (1)	4 (0)	0	2
PON3	2 (1)	1 (0)	1	4
Total	31 (7)	20 (3)	3 (2)	14

Table 3.3. Experimental comparison. [a]Experimental unpublished data; the experimental validation is done by resequencing. [b]number of SNPs found; [c]number of experimental SNPs found.

in di□erent isoforms from the same gene are kept separated for two reasons, biological and technical:

1) the meaning of the same mutation in di□erent isoforms expressed within a distinct spatial or temporal pattern could lead to di□erent biological e□ect;

2) the design of probes or primers for sequencing purpose must take into account the genomic neighborhood (introns, untranslated regions, promoters) that distinguish one isoform to one other.

All the 12,159+1,583 filtered SNPs were aligned on the genomic sequences and for 6,393+733 of them the procedure returned a probe that fit the technical specifications required to design the Golden Gate. After the whole process 690 new genes, potentially related to obesity, were disposable to further investigation.

3.2.6 In silico comparison

To test the performance of the SNP search procedure we compared our results with HaploSNPer (with default parameters), a web tools that was, at the time in which we developed the tool, the only released pipeline for SNP detection able to perform SNPs detection in a set of template sequences.

We used a random sample from OHR, composed by 138 genes[3]. 67 out of 138 (48.8%) genes appear to contain SNPs for at least one pipeline (our or HaploSNPer). We detected 255 SNPs in 52 genes and HaploSNPer 1,338 SNPs out of 56 genes. No overlapping between genes. On average our pipeline and HaploSNPer detect 3.81 and 19.97 SNPs per gene. 1,338 SNPs in 56 genes highlights problems in the SNPs selection procedure.

A deeper data analysis shows that this high mutation rate discovery is due to problems with paralogs and orthologs sequences discrimination. We found that for at least then genes HaploSNPer detects more than 10 SNPs in a short stretch of DNA (few tens of bases). For two of these genes, *IGF2* and *ACP1*, HaploSNPer detect more than 100 SNPs on a single sequence. This problem appears when sequences deriving from di□erent organisms are aligned (and this is the case) [RSS01]. A second source of error that a□ect the procedure adopted by HaploSNPer is the use of the templates as a source of variation (to improve the sensibility of the method). In an inter-species alignment, the template sequence must be removed from the alignment in order to avoid false variation discovery (hybrid-SNPs). In 5 genes in which Haplo detects SNPs (*KLF5, GHRHR, IKBKB, INPPL1* and *LACTB*) the variation originate if the polymorphism appear in the seed sequence. When genes with paralogs or with hybrid-SNPs were removed, the performance in term of average SNPs per genes appear to be similar (2.98 and 2.21 SNPs

[3]We have reduced the dimension of the test set to 138 genes because the analysis of the results was performed by hand

Gene symbol	P	SNP type	SNP location	Chromosome	Source
TBC1D1	0.000137	G>A	Intron 2	8	resequencing
CALR	0.000601	A>G	Exon 1	2	in silico
PPARG	0.000690	A>G	5'UTR	13	literature
JAK3	0.000736	T>C	Exon 23	2	in silico
NT5E	0.000937	T>C	Exon 7	1	in silico
RRAGD	0.002064	T>C	Intron 3	1	literature
MC4R	0.002220	A>G	Exon 1	1	literature
VCAM1	0.002281	T>C	Exon 5	4	in silico

Table 3.4. List of genes associated with BFT with P<0.10.

on average) for our pipeline and HaploSNPer. The Number of total SNPs found are 228 and only 16 (7%) are common.

Then we compared the performance on a set of 10 experimental validate genes with 14 validate SNPs. As is shown in table our pipeline detect 31 SNP and HaploSNPer detect 20 SNP wheres only 3 SNPs are shared. More interesting are results on the experimental validated SNP where our pipeline detect 7 SNP (50%) and HaploSNPer 3 (21.5%) showing a consistent overperformance of our selection strategy.

3.3 Probes testing

Sequencing of fragments amplified from 60 obesity related genes identified 84 SNPs, 67 of which were chosen for genotyping. Other 131 SNPs obtained from published literature on markers associated with fat deposition traits in pigs were included in the genotyping SNP panel. In silico expressed sequence tags database mining reported $733 + 6,393 = 7,126$ SNPs. From the 7,126 SNPs found in silico 565 were analyzed whereas 198 were selected from published literature on markers associated with fat deposition traits in pigs. On the whole adding the different sources from which we identified SNPs, 763 SNPs were included in the GoldenGate panel that was used to perform the association study.

This study was conducted on two extreme and divergent groups of two generation unrelated females of these triplets (evaluated in the period 1996-2007) chosen according to their Estimated Breeding Values (EBV) for BFT (280 with most negative and 280 with most positive EBV) within a population of ~12,000 pigs. Average EBVs for the most negative and positive selected pigs were –9.4 ± 1.6 and +8.0 ± 5.95 mm, respectively. Genomic DNA was extracted from blood using standard protocols.

For 86 SNPs the designed GoldenGate tests were not successful leaving 677 SNPs: 169 SNPs were monomorphic in the analyzed pigs; 187 SNPs had a MAF<0.05; and 321 had MAF >0.05.

Allele frequency differences between the two extreme groups of pigs chosen according to their BFT EBV were evaluated for 321 SNPs with MAF >0.05. Of these SNPs, 65 showed P nominal value <0.10. Thresholds of 0.10 and 0.20 corresponded to P nominal values of 0.003 and 0.025, and included 8 and 30 SNPs, respectively. For brevity, only genes whose polymorphisms were significantly associated with BFT at P<0.10 are reported in Table 3.4. . The role of the other significant genes on BFT remains to be further evaluated.

Most of these genes map in chromosome regions in which QTL for BFT and other related traits have been localized. The recruitment strategy of candidate genes based on the human interactome followed by an association study of in silico detected SNPs represents an innovative approach that may open new possibilities for the identification of genes affecting obesity related traits in pig enforcing the role of this species as a model for human obesity.

Chromosome	*NumberSNPsMapped*[a]	*Frequency SNP*[a]	*NumberSNPsMapped*[b]	*Frequency SNP*[b]
chr1	6055	2.04E-005	6299	2.13E-005
chr10	1622	2.01E-005	1487	2.23E-005
chr11	1726	2.05E-005	1732	2.17E-005
chr12	1403	2.18E-005	1355	2.36E-005
chr13	3836	1.80E-005	3504	2.41E-005
chr14	3600	2.34E-005	3900	2.63E-005
chr15	2631	1.72E-005	2516	1.87E-005
chr16	1702	1.98E-005	1663	2.15E-005
chr17	1548	2.23E-005	1546	2.40E-005
chr18	1200	2.00E-005	1203	2.22E-005
chr2	3138	1.92E-005	3026	2.16E-005
chr3	2566	1.80E-005	2479	2.01E-005
chr4	3374	2.44E-005	3529	2.59E-005
chr5	2153	1.95E-005	2193	2.18E-005
chr6	2961	1.89E-005	2669	2.16E-005
chr7	3126	2.33E-005	3324	2.44E-005
chr8	2571	1.74E-005	2406	2.01E-005
chr9	2946	1.94E-005	2875	2.17E-005
X	1384	1.03E-005	1293	1.03E-005
Y	10	6.24E-006	0	-
Total	49552		48999	

Table 3.5. Probes distribution on chromosomes in [a] $susScrofa10$ and [b] $susScrofa9.2$.

3.4 Illumina's PorcineSNP60 BeadChip array annotation

High density single nucleotide polymorphism (SNP) genotyping arrays have been recently developed in several farm animal species. Their use makes it feasible genome wide studies for different purposes, including also new applications like genomic selection. However, to fully exploit the potential of these arrays, SNPs, that are usually anonymous, should be characterized and precisely assigned to their chromosomal position. These two issues are particularly challenging when the genome sequences are still under refinement, as in the case of most farm animal species. In particular, for the pig several versions of the genome have been preliminarily released to the public. Recent updates provided new versions of the porcine genome (Sscrofa9 [ABC+10], `http://www.ensembl.org/Sus_scrofa/Info/Index` August 2010; and Sscrofa10, Sept. 2010). The Illumina's PorcineSNP60 BeadChip array [RCA+09] includes 62,163 SNPs, that were originally positioned on the Sscrofa7 genome pre-assembly version. The last effort made on this project was to assign these SNPs to the two latest versions of the pig genome as a first step for their functional annotation.

3.4.1 SNP assignment and annotation

Probes of the chip were aligned to Sscrofa9 and Sscrofa10 using MEGABLAST[4]. Only matches with identity $\geq$ to 94% were considered. Probes with unique match on the genome were considered as mapped. In case of multiple significant matches, the alignment with the highest score was considered. The position of each SNP was calculated on the basis of the alignment data knowing that the polymorphic site is the first base that comes after the last base of the probe. We used this simple formula: $SNP_p = Gen_p \pm (51 - Probe_p)$, where SNP_p is the SNP position, Gen_p is the position of the last base aligned in the reference genome and $Probe_p$ is the position of the last base aligned in the probe. The sign of the operation is con-

[4] parameters: `-e 1e-10, -W 28, -p 90` no filtering for low complexity.

SNP type	*Number*
Intergenic	31892
Intron	14347
Downstream	2404
Upstream	2377
3Prime	103
5Prime	36
Synonymous	699
Non-Synonymous	291
Stop Gain	5
Start Gain	4
Stop Lost	0
Start Lost	1
Splice site	25

Table 3.6. SNP effect on genes. We used the annotation provided for the susScrofa9.2 assembly.

cordant with the alignment strand (plus or minus). For each SNP position we searched the reference genome for discordance with the SNP reported in the Illumina definition file. We detect a discordance if the base in the SNP position was di□erent from the two bases called in the SNP file. According to these criteria, 48,999 and 49,552 probes were uniquely assigned to Sscrofa9.2 and Sscrofa10, respectively (Table 3.5 reports distribution of SNPs per chromosome). Transcript records for 20,460 genes were retrieved from Sscrofa9.2 using BioMart (`http://www.ensembl.org/biomart/index.html`). The probes defined as mapped were processed using SNPEFF (`http://snpeff.sourceforge.net/`) to predict their e□ects (Table3.6). Most of the annotated SNPs were intergenic (65.63%) whereas only a small fraction were predicted as non-synonymous (291) and a few others (35) might have a direct functional e□ect (Table 3.6). Two files include the list of SNPs uniquely assigned and annotated for the Sscrofa9.2 and Scrofa10 genome versions. These files will be provided to the scientific community after the publication of the results.

Chapter 4

Searching for Emergent Properties of Cellular Systems

4.1 Observing gene activity

Currently, it is possible to observe the activity (over- and under- expression, presence or absence of mutations) of almost all molecules of a given type (mRNA, miRNA, DNA) in a single screen using high-density chips [GN08], or sequencing related techniques [HJ08, WGS09]. Lately, the number of studies using microarray platforms for analysis of mRNA are quickly being followed by similar analyses related to miRNAs [VCL+06, YKV+08]. Only recently both types of variables were analyzed simultaneously [LPP+07, LFG+07, PBK+10], while, typically, both types of data are analyzed in search for (i) molecules sharing similarity, using simply the expression available at the time (*unsupervised* approaches, [But02]) e.g. clustering [Qua01, MO04] and association networks [MNB+04, MLB08, NRT+07] or (ii) similarity with -or dependency from- other types of traits, providing for example clinical classes or other non-molecular information on the samples (*supervised* approaches, [But02]) i.e. Significant Analysis of Microarray (SAM [TTC01]), Gene Set Enrichment Analysis (GSEA [MLE+03]).

However, this approach implies to analyze separately different aspects of a system (e.g., transcriptional and/or post-transcriptional mechanisms) and the results may not be concordant with analyses of the system as a whole. For example, interactions among miRNAs and mRNAs may be underestimated or completely overlooked. This lack of information can be expressed as missing the *emergent* properties of the system. While the concept of emergent properties is well known in Systems Theory, it has only recently become an important concept in the area of life sciences, thanks to the relatively new approach of Systems Biology [Hoc05, ATPP06b, ATPP06a]. Emergent properties arise from hierarchical integration of the individual components and organizational levels of complex systems, and, biologically, they are only manifest when the organism is considered in its entirety.

Analogous to emergent properties in systems biology is the concept of latent variables in multivariate statistics. Latent variables are so-called hidden variables generated in certain types of multivariate analysis (e.g. factor analysis, see below) which are not evident in original observed data. Rather, these latent variables emerge from consideration of the covariance patterns when a large number of relevant variables are analyzed simultaneously. These latent variables may reflect a summarization of causal indicators underlying observed biological variability. Given the parallelism between biological system's emergent properties and latent variables, we sought- quite naturally- to investigate the ability of latent variables to describe emergent properties, by applying multivariate analysis simultaneously to different parts of a biological system, and notably to transcriptional and post-transcriptional data.

Previously, successful parallel multi-platform analyses were performed integrating genomic and transcriptional level, by using CGH arrays or SNPs and cDNA arrays [YWF+06, LTTD+00]. This approach portend to explain variations observed at the transcriptional level, based on information at the genomic level. These approaches can annotate and map different types of probe IDs onto genomic coordinates [YCL+06], or add analyses at the translational level [MHH+05].

Model	Tumor Grade	Anaplastic	Glioblastoma	Gliosarcoma
SVM	(**0.92**, *0.045*)	(**1**, *0.0015*)	(**0.83**, *0.08*)	(**0.83**, *0.045*)
NB	(**0.92**, *0.045*)	(**1**, *0.0015*)	(**0.58**, *1*)	(**0.83**, *0.077*)
NN	(**0.92**, *0.045*)	(**1**, *0.0015*)	(**0.83**, *0.06*)	(**0.83**, *0.045*)
kNN	(**0.92**, *0.045*)	(**1**, *0.0015*)	(**0.83**, *0.08*)	(**0.92**, *0.018*)

Table 4.1. Model Selection, Alternative Classification Analysis. Tumors type and grade dual discrimination. In bold **Accuracy**; in italic *p-value*. SVM: Support Vector Machine; NB: Naïve Bayes; NN: Neural Network; kNN: k-Nearest-Neighbors.

However, to date, simultaneous analysis of miRNA and mRNA from the same tissue have used only profile correlations [LPP+07].

Herein, we expanded analyses of molecular covariation beyond correlation of expression profiles by using the multivariate statistical procedure of multiple or common Factor Analysis (FA, [JW02], see Section 2.1.4). FA was used extensively to cluster microarray data [She01, Pet02b, LSB+05b]. The use of the *a priori* knowledge on how each sample maps on tumor classes to constrain the relation among latent variables under study and the factors obtained permits further data interpretation. In other words we perform a FA that is driven by data (hypothesis) pre-established to find latent variables that could be investigated to obtain biological information [CGP+06b].

To constrain the factor model we used Linear Discriminant Analysis (LDA, [JW02]), a technique used to classify a set of observations into categories (a dichotomy in our case). Besides LDA, other classifiers (Support Vector Machine, Naïve Bayes, Neural Network and k-Nearest-Neighbors) were also tested and performances are listed in Table 4.1. We only briefly mention here that most of the performances are identical for all the classifiers, and only for the Glioblastomas discrimination LDA shows slightly more accuracy. These results indicate that the classification analysis is robust and gives stable results independently from the choice of the classification algorithm.

Factor analysis proceeds from a matrix of pair-wise correlations to extract a small number of *factors* that describe the major patterns of common covariation. More formally, the common factor model is based on the equation $D = LF + E$, where D are the observed variables, L are the common factors, F are the coefficients or scores of the factors and E are the unique factors, under the assumptions that the unique factors are uncorrelated with each other and that F and E are independent. Since only common variation is analyzed, these individual factors describe the latent structure underlying the major patterns of molecular covariation.

The sign and magnitude of the factors coefficients reflect the extent and direction of the correlation between each variable and individual factor and describe the relative contribution of each variable to a particular pattern of multivariate changes.

FA derives a set of *factor scores* that gives the relative location of each item in the reduced latent variable subspace. The resultant factors, coefficients and scores are interpreted in light of biological knowledge about the specific data under study. FA can define a biological model about the underlying nature of molecular covariation (e.g. number of patterns of covarying elements and their relative importance). These models are evaluated both biologically and statistically and subsequently used to explain the structure and dynamics of complex biological systems.

There exist several approaches to perform data reduction and classification (see for example Bayesian classifiers [LIT92, Fri98, PKS+07], Support Vector Machine [FCD+00], K-nearest-neighbor [TCRR+02]), however, FA has already been used successfully in various applications related to molecular biology, like the identification of multidimensional patterns of molecular covariation able to describe protein's structures [AZFD05].

More classical approaches have been designed for effective clustering in the analysis of cDNA microarrays and Expressed Sequences Tag (ESTs) [Pet02a], as well as in specific applications to identify genes and pathways related to biological categories that could be associated to relevant phenotypes in both yeast and humans [LSB+05a] or to test and validate hypotheses on the

Model	Tumor Grade	Anaplastic	Glioblastoma	Gliosarcoma
1	-	-	-	-
2	F 2 (**0.92**, *0.045*)	F 2 (**1**, *0.015*)	-	-
3	F 2 (**0.92**, *0.045*)	F 2 (**1**, *0.015*)	F 1+F 2 (**0.92**, *0.015*)	F 1+F 3 (**0.92**, *0.018*)
4	F 2 (**0.92**, *0.045*)	F 2 (**1**, *0.0015*)	F 1+F 2 (**0.92**, *0.015*)	F 1 (**0.92**, *0.018*)
5	F 1 (**0.92**, *0.045*)	F 1 (**1**, *0.0015*)	-	F 5 (**0.92**, *0.018*)

Table 4.2. Model selection using discriminant analysis. Tumors type and grade dual discrimination. In bold **Accuracy**; in italic *p-value*

association of gene expression to cisplatin resistance in ovarian cancer cell lines [CGP+06a]. One of the advantages of this approach over hierarchical clustering is the possibility to include genes in more than one category. More recently, FA was used to filter informative and non-informative data from microarray for gene expression [KLVS+10]. Variations of classical FA (Bayesian factor analysis) have been used to identify the latent structure that describes the relationship between transcription factors and genes, using microarray data [PW07]. Previously, this approach was used to perform gene network reconstruction in *E.Coli* taking advantage of literature information, DNA sequences and expression arrays [SJ06].

This chapter describe an application of FA to the composite analysis of multilevel molecular data.

4.2 Multilevel latent structure

Because miRNAs and mRNAs are processed together, from now on, Factors will always be likely to include both mRNAs and miRNAs in their composition. To avoid confusion on the meaning of the word *gene*, we use the term *coding genes* to refer to mRNAs and the generic term *genes* to refer both to mRNAs and miRNAs. The interpretation of factors based on associating them to mRNAs/miRNAs (separately considering positive/negative scores) is a novelty of the presented approach, and will be discussed in details in the coming sections. In particular, in the following we will describe how to identify and interpreted the latent factors, both using mRNA and miRNA (indirect) functionalities. Then, we will describe the biological structure emerging from this analysis, speculating on its clinical meaning. Finally, we will o□er a comparison with the results of an analysis done in parallel.

4.2.1 Identification of *multilevel* latent structures

We performed several Factor Analyses obtaining Models characterized by 1 to 5 factors (named here Model n, $n = 1,..,5$). Kaiser criterion [vZ85] has been used to identify the number of factors that show a large variance (common variance in each factor greater than a given threshold, t) and therefore carry a large amount of the information hidden in the data. Given $t = 1$ the number of information-rich factors appears to be 4. Therefore, FA was performed with a growing number of such factors, from the one with higher variance, up to 5, to test the appropriateness of the variance threshold. We then confirmed the validity of a subset of the Models using LDA to identify which factor (or linear combination of factors) was able to best classify tumor grade and histopathology, based on the statistical significance of Fisher exact test [SR03]. This test, suited for contingency tables where one or more expected frequencies are below 5, evaluates the null hypothesis associated with LDA that there are no statistically significant di□erences between the *a priori* clinically defined groups. The models for which the null hypothesis was rejected were retained (see Table 4.2 and Methods for details) . Therefore, we performed 4 LDA, namely between a class and its complement: i.e. high/low grade, anaplastic/non-anaplastic, glioblastoma/non-glioblastoma and gliosarcoma/non-gliosarcoma, following the original classification in [LPP+07]. We did not consider oligodendroglioma relevant, because of a single sample available.

Factor	Ontology Terms	Ontology
$F1^+$	Response to external stimulus	GO_BP
	Secreted, glycoprotein	SP
	Plasma Membrane, transucer, extracellular, receptor	GO_MF,GO_CC, SP
$F2^-$	Signal, glycoprotein	SP
	Cell Adhesion	SP
	Extracellular region	GO_CC
$F3^+$	Gene Expression	GO_BP

Table 4.3. Functional analysis of the factors in Model 3. GO: Gene Ontology; BP: Biological Process; CC: Cellular Component; MF: Molecular Function; SP: Swiss Prot.

Model 3 appears to be the most suitable, since it is able to discriminate between anaplastic and non-anaplastic tumors with 100% accuracy (based only on Factor 2) and the other two types of tumors with $\simeq$ 92% accuracy. Since anaplastic tumors are low grade tumors, Factor 2 is relevant in the identification of low grade tumors in general with $\simeq$ 92% accuracy, since the only oligodendroglioma appears to be elusive. It is worth noting that Model 4 shows the same performance scores, but with a greater number of factors and Factor 4 does not appear to be involved in class identification.

4.2.2 Interpretation of *multilevel* latent structures

4.2.2.1 mRNA functional analysis Working solely on Model 3, the mRNAs in each factor were processed to detect enriched Gene Ontology (GO, [Con01]) terms or UniProt (SP, [BAW+05]) keywords. The magnitude and sign of the factor scores (not the factor coe□cients from the eigenvectors) give their relative relationship with the expression of miRNA and mRNA. Consequently, each row in the 3 factors score matrix ($F1$, $F2$ and $F3$) was split into positive and negative portions ($F1^+$ and $F1^-$; $F2^+$ and $F2^-$; $F3^+$ and $F3^-$) and analyzed separately. $F1^+$ is associated with GO terms related to response to stress and external stimuli. Terms from SP keywords like *secreted* and *glycoprotein* were also found in this subset. Thus this factor appears then to be related with cell functions that process signal from the external environment to the cell with membrane receptors involved to the signal transduction. $F2^-$ is also involved in the signaling, including categories related to cell adhesion, it appears then to be related to functions like chemotaxis that are involved in inflammation processes.

Finally, $F3^+$ contains coding genes that are related to the biological process that goes under the general term *gene expression*. Gene expression includes all the mechanisms such as transcription, translation, RNA maturation, proteins transport and ubiquitination by which information coded in the DNA is converted to a functional product. All results are summarized in Table 4.3.

4.2.2.2 miRNA *Indirect* functional analysis Since miRNAs are not included in any ontology database, we performed an indirect functional analysis by screening the functional terms associated with the experimentally validated target coding genes of the miRNAs, extracted from TarBase [PRS+09]. Once the target coding genes were identified, they were manually annotated via GO terms or SP keywords, as above (see Table 4.4).

4.2.2.3 mRNA/miRNA complex functional annotation We then checked the functional classification's coherence between the indirect and direct functional analysis, within each significantly annotated factor (i.e. $F2^-$, $F3^+$, since no miRNA appeared in $F1^+$). Thus, globally speaking, $F1^+$ annotation is unchanged and related to functions that are responsible for signal transduction. In $F2^-$, 3 out of 7 target coding genes (CXCL12, TM6SF1 and AGTR1) are annotated with terms that can be associated to the categories significantly varied in the mRNA functional analysis: $F2^-$ is then confirmed to be a factor involved in functions related with adhesion and/or chemotaxis. For the miRNAs in $F3^+$, 5 out of 8 target coding genes (ARID4B,

miRNA	Gene Name	Gene Description
	AOF2	amine oxidase (flavin containing) domain 2
	ASXL1	kiaa0978 protein
	C19ORF2	chromosome 19 open reading frame 2
	CHD7	chromodomain helicase dna binding protein 7
	CHD8	chromodomain helicase dna binding protein 8
	ETV1	ets variant gene 1
	GTF2H3	general transcription factor iih, polypeptide 3, 34kda
	HEY1	hairy/enhancer-of-split related with yrpw motif 1
	HIC2	hypermethylated in cancer 2
	ID4	inhibitor of dna binding 4, dominant negative helix-loop-helix protein
	KCNH2	potassium voltage-gated channel, subfamily h (eag-related), member 2
	MAZ	myc-associated zinc finger protein (purine-binding transcription factor)
hsa-miR-17-5p	NONO	non-pou domain containing, octamer-binding
hsa-miR-20b	OLIG2	oligodendrocyte lineage transcription factor 2
	POLR2E	polymerase (rna) ii (dna directed) polypeptide e, 25kda
	PPP2R1A	protein phosphatase 2 (formerly 2a), regulatory subunit a (pr 65), alpha isoform
	PRMT5	protein arginine methyltransferase 5
	RBBP4	retinoblastoma binding protein 4
	RBPJ	recombining binding protein suppressor of hairless (drosophila)
	RERE	arginine-glutamic acid dipeptide (re) repeats
	RTF1	rtf1, paf1/rna polymerase ii complex component, homolog (s. cerevisiae)
	SOX12	sry (sex determining region y)-box 12
	SSRP1	structure specific recognition protein 1
	STAT5B	signal transducer and activator of transcription 5b
	TBL1XR1	transducin (beta)-like 1x-linked receptor 1
	TCF12	transcription factor 12 (htf4, helix-loop-helix transcription factors 4)
	TCF3	transcription factor 3 (e2a immunoglobulin enhancer binding factors e12/e47)
	TCF7L1	transcription factor 7-like 1 (t-cell specific, hmg-box)
	YBX1	y box binding protein 1
	ZNF195	zinc finger protein 195
	ZNF43	zinc finger protein 43 (htf6)

Table 4.4. $F3^+$: miRNA and mRNA annotated with Transcription Regulation term. Number of mRNA found in Factors after multilevel analysis.

MYLIP, HIPK3, E2F1 and NCOA3) are functionally related with the *gene expression* term found in the mRNA functional analysis. Interestingly, most of the terms (4/5) are related with mechanisms of transcription regulation and only one with protein ubiquitination.

After direct and indirect annotation, 2 miRNAs and 31 human coding genes in $F3$ were selected as belonging to the same category (see Table 5). Not surprisingly, most of the coding genes in this list are not predicted to be targets of the 2 miRNAs that appear in the factor. In fact, the biological meaning of the result is a set of genetic elements that share covariability in the expression pattern and we know that, e.g. in animals, most of the control on gene expression is performed by tuning translation. Therefore, the levels of miRNAs and the mRNAs of direct targets are not directly correlated. As it is also suggested in [LPP+07] we can imagine that our list of coding genes contains the possible subset of indirect targets (functionally related with the regulation of the transcription) of two miRNAs: miR-17-5p, and miR-20b. Globally, $F3^+$ is confirmed to be associated with gene expression, with transcription regulation being the most common mechanism of expression.

4.2.3 Emergent properties

Since the transcription regulation term ($F3^+$) appears to give the clearest biological information, coherent in mRNAs and miRNA, we focused our e□orts on this part of the analysis.

Latent structure chromosomal localization Most of the miRNAs in $F3^+$ belong to two polycistronic miRNA genes where miRNAs are lying in close proximity on the chromosome. (The named *clusters* are given in italics throughout the paper to improve readability and avoid confusion with clusters emerging from supervised or functional analyses). These polycistronic miRNA genes are involved in cell proliferation, apoptosis suppression, tumor angiogenesis [Men08] and T cell leukemia [LLLR07]. The first polycistronic gene (miR-17-92) is composed by 7 miRNAs and maps on Chromosome 13 whereas the second one (miR-106-363) maps on Chromosome X

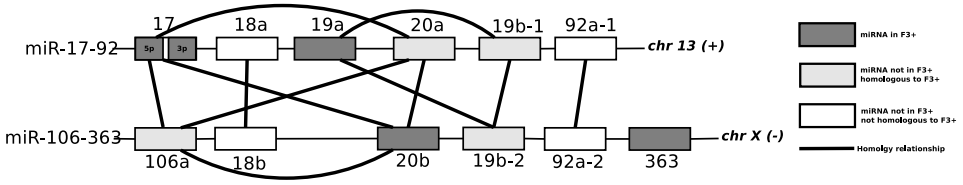


Fig. 4.1. Organization of miRNA clusters *miR-17-92* and *miR-106-363*. Structure of the two polycistronic miRNA gene and the relations between miRNAs.

and contains 6 miRNAs, details are shown in Figure 4.1. The two *clusters* are closely related, in fact, each miRNA on one *cluster* has at least one homologous in the other *cluster* except for miR-17-3p and miR-363 that do not share homology with the other miRNAs (shown in Figure 4.1).

As further corroborating test, we observed that, when searching the target coding genes of homologous miRNAs (miR-20a, miR-17-5p and miR-106a) the list of predicted targets (Targetscan, [LBB05]) is identical for all miRNAs. Moreover, we notice that only two homologous groups of miRNAs in the *cluster* (miR-18 and miR-92) are not part of $F3^+$. If we look at their sequence in detail we observe that they are very similar to miR-20a with only two mismatches: one in the loop (miR-18a and 18b) and one after the supplementary pairing region (miR-18b). This can represent a partial functional redundancy since all the known key regions in target recognition are identical. Conversely, miR-92 does not share any significant homology with the other members of the *cluster* (except for the seed region with miR-363).

Taking into consideration all the redundancies in the *clusters*, most of the transcript targets in $F3^+$ are probably under the regulation effect of the expressed miRNAs. It is worth noting that a cross-hybridization effect in miRNAs could be considered the mechanism responsible for these association in *clusters*. But, as reported by the authors of the dataset [LPP+07], each primer and probe contained zip-coded sequences specifically assigned to each miRNA to increase the specificity of each reaction so that even small differences in miRNA were amplified and detected. So, this artifact can be discarded as explanation for the emerging of *clusters* of miRNA.

Statistical relevance Interestingly, in $F3^+$, only 2 miRNAs (hsa-mir-9 and hsa-mir-130b) out of 7 do *not* belong to any of these two *clusters*. Their role was shown respectively to be related to the molecular pathogenesis of ovarian cancer [LOF+08] as well as to schizophrenia and Human T-cell leukemia Virus-1 (HTLV-1) transformation [BGA+07, YiYB+08]. Six more miRNAs (miR-106a, miR-18a and miR-18b, miR-20a, miR19b-1 and miR-19b-2) that belong to these two *clusters* could not be part of our analysis, as they were not part of Liu's original dataset. Given the high density of miRNAs in these *clusters*, we used the hypergeometric distribution to compute the probability associated with the hypothesis that a random sampling would give the same result in terms of number of *cluster* members in *cluster* miR-17-92 (3 members out of 4 total), in *cluster* miR-106-363 (2 members out of 3 total) and in both (5 members out of 7 total). The reference group for computing the probability consists of the total number of detected miRNAs (93). The resultant probabilities were Bonferroni corrected and were equal to 3.6×10^{-3}, 0.045 and 2.3×10^{-7} respectively. All three are statistically significant.

4.2.3.1 Speculations on molecular-clinical implications Ultimately, we speculated on how the two *clusters* between gliosarcomas and non-gliosarcomas. This choice is due to the fact that our analysis has shown that the combination of factors that carry the more coherent functional information (both from miRNAs and mRNAs signals) was the combination able to discriminate glioscarcomas from other tumors. Believing that such a coherence could hide strong biological meanings we focused on gliosarcomas the efforts to detect emergent properties. This complex task, that cannot be fully explained with the data and results in hand, can take

(a) Tumor Grade

	High/Low Grade	
	P High	**P Low**
High	**9**	**0**
Low	**1**	**2**
	p=0.045	

(b) Anaplastic

	Anaplastic	
	P Anap	**P *Anap**
Anap	**10**	**0**
* Anap	**0**	**2**
	p=0.015	

(c) Glioblastoma

	Glioblastoma	
	P Glio	**P *GLio**
Glio	**5**	**1**
* Glio	**1**	**5**
	p=0.08	

(d) Gliosarcoma

	Gliosarcoma	
	P Gsar	**P * Gsar**
Gsar	**2**	**1**
* Gsar	**2**	**7**
	p=0.23	

Table 4.5. Performances of Model 3 using only miRNA data. These Tables shows the classification performances of Model 3 on expression data of miRNA only. Significant classifications in bold ($p < 0.05$). Anap: Anaplastic; *Anap: non Anaplastic, Glio: glioblastoma; *Glio: non glioblastoma, Gsar: gliosarcoma; *Gsar: non gliosarcoma.

advantage of intriguing observations emerging from the analysis. We notice, in fact, that the presence of the sarcomatous element, that derives from an endothelial hyperplasic lesion [FG55], is a characteristic of these kinds of tumor.

The hyperplasic lesion is a proliferation of vessel-wall components that contains endothelial cells, myofibroblast, smooth muscle cells and other components of the vascular endothelium [KTF+86]. In [Men08] it is also shown that *cluster* miR-17-92 is related to solid tumors angiogenesis. The finding of this *cluster*, and the homologous miR-106-363, in the factor that contributes to discriminate gliosarcomas, could then indicate an involvement in the development of the sarcomatous element.

4.2.4 Identification and interpretation of *Simple* latent structures

In this Section we present results obtained from analyzing with FA and LDA the two datasets (mRNA and miRNA) separately. Our original hypothesis dealt with the ability of the complex analysis to identify emergent properties. To evaluate this hypothesis we produced a 3 factor model with factor analysis on the two expression matrices separately. Next, we analyzed the two series of factor scores using separate LDA. In this Section we identify with $F_{mi}i$ Factor i obtained from the miRNA dataset and with F_mj Factor j from the mRNA dataset (Fk continues to identify Factor k from the joint dataset. Regarding the identification of the latent structures, as expected and given the larger size of the mRNA matrix, the results in terms of discrimination power among tumor classes and the functional analysis are unchanged. However, the situation is di□erent for the miRNA data.

As shown in Table 4.5 only high/low grade tumors and anaplastic/non analplastic categories are predicted with the same accuracy (and on the same factor, $F_{mi}2$). The accuracy is lower, 0.83 (p = 0.08) versus 0.92 (p = 0.015) for the glioblastoma/non-glioblastoma category. This occurs because one of the glioblastomas is predicted as a non-glioblastoma. Furthermore, the discrimination appears to be based on a linear model composed only by $F_{mi}1$ and not on a combination (see $F1$ and $F2$ in the complex analysis). The discrimination between gliosarcomas and its dual class is the worst, as accuracy drops to 0.75 (p = 0.23) and $F_{mi}3$ is not used in discrimination.

For what concerns the interpretation of the latent structures, out of the 18 miRNAs selected, 9 are in common with the joint analysis and 9 represent a new set of miRNAs. Five of the miRNAs in the new set are associated with biological terms, and only one (hsa-miR-126) is shared by more than one factor ($F_{mi}1$ and $F_{mi}2$). $F_{mi}1$ contains 5 terms, $F_{mi}2$ 2 terms (a subset of $F_{mi}1$) and $F_{mi}3$ 2 terms (Table 4.6). These are related with the regulation of the transcription (in $F_{mi}1$ and $F_{mi}3$) and they show some overlap with the mRNAs Factors annotation. Namely, biological terms in $F_{mi}1$ overlap with all the three F_m whereas terms in $F_{mi}2$ overlap only with

miRNA	Annotation	Factor
hsa-miR-106b	response to gamma radiation	$F_{mi}1$
hsa-miR-29b	cell fate determination, positive regulation of gene expression	$F_{mi}1$
hsa-miR-23a	*cell adhesion, chemotaxis, positive regulation of monocyte chemotaxis, extracellular space*	$F_{mi}2$
hsa-miR-126	membrane to membrane docking, cell junction	$F_{mi}1$, $F_{mi}2$
hsa-miR-155	*regulation of natriuresis, regulation of cell growth positive regulation of inflammatory response, regulation of blood vessel size by renin-angiotensin, Transcription regulation, Nucleus negative regulation of cell proliferation, negative regulation of cell proliferation response to toxin, response to UV, positive regulation of programmed cell death cyclin-dependent protein kinase inhibitor activity nuclear matrix RNA binding, protein binding*	$F_{mi}2$, $F_{mi}3$
hsa-miR-210	cell-cell signaling	$F_{mi}3$
hsa-miR-27a	regulation of transcription, DNA-dependent	$F_{mi}3$

Table 4.6. miRNA from Simple Structure Analysis. Functional Annotation. Only miRNAs that give an annotation are listed in the table. $F_{mi}1$: Factor 1; $F_{mi}2$: Factor 2; $F_{mi}3$: Factor 3. *Italics*: miRNAs and annotations shared with the Complex Analysis.

F_m2. Terms in $F_{mi}3$ are found both in F_m2 and F_m3. With respect to the comparison to the complex analysis, since these miRNAs are mostly clustered in homologous factors it is possible to associate $F_{mi}3$ with $F1$, $F_{mi}2$ with $F2$ and $F_{mi}3$ with $F1$). The miRNAs shared with the *complex* analysis and that return an annotation are in $F_{mi}2$ (both miR-155 and miR-23a) and $F_{mi}3$ (miR-155).

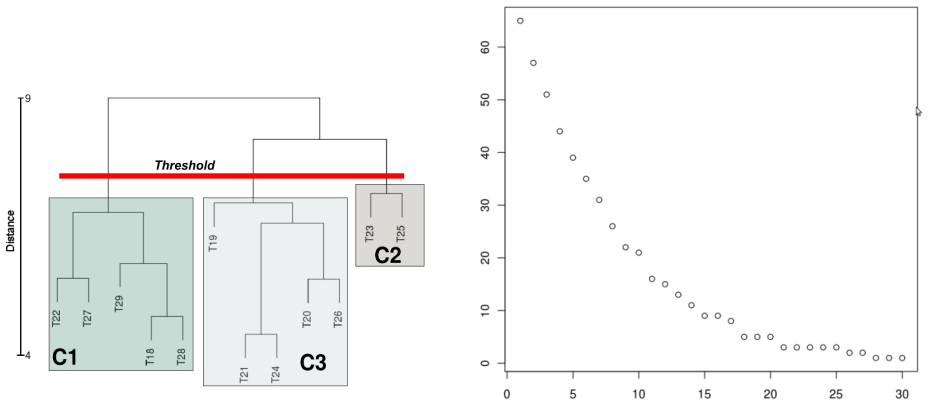
However, without the joint analysis there is no obvious rationale to associate miRNA factors with mRNA factors. This is because, crucially, the 18 miRNAs obtained are distributed over factors that are decoupled from the factors returned from the *simple* mRNA data analysis. Therefore this approach does not suggest any obvious association between the two sets of factors. As a consequence, the interpretation of this latter (*simple*) analysis is limited to the indirect functional annotation of this small set of miRNA (Table 4.6). Therefore, the activation of the polycistronic clusters miR-17-92 and miR-106-363 does not emerge when miRNAs are analysed separately.

In summary, combining the two datasets and applying FA and LDA, provides an obvious way to associate the translational and post-translational information. In particular, although the mRNA latent structure is the same in the simple and complex analysis, and consequently the functional annotation is the same, hidden signals present in the smaller dataset (miRNA set) appear to be amplified by the signals present in the larger dataset (mRNA set) thanks to their association in a common latent structure.

4.2.5 Alternative approaches

In order to compare our approach to other likely tools, we choose to analyze the mRNA/miRNA dataset with 3 methods. First we used hierarchical clustering [ESBB98] an unsupervised and extremely popular tool for microarray analysis. Given the poor results obtained, we sought to preprocess the dataset with SAM, another very popular, but supervised method that imposes a structure in the data according to an *a priori* classification of the dataset (i.e. grade and tumor type), to favor the identification of elements related to the clinical classification.

Clustering has been successfully used in the past to classify clinical samples based on microarray data, therefore the current analysis aims at showing more the inability of this approach to identify emerging properties (in this case the polycistronic miRNA clusters) rather than its limited ability to classify samples.



(a) Sample Clusters (b) Clusters size

Fig. 4.2. Results of Clustering. Figure 4.2(a) shows the sample clusters obtained without preliminary SAM analysis. Results are identical for the analysis performed after SAM analysis. Figure 4.2(b) shows the variability in the clusters number around threshold values varied in the interval [0.4-3] for SAM + Clustering analysis.

4.2.5.1 Cluster Analysis R was used to perform clustering on the whole dataset based on euclidean distance and correlation between each couple of experiments, obtaining two sets of $n(n-1)/2-n$ = 120 distances. Sample clusters identified two low grade tumours (T21 and T24) clustered together whereas the third low grade oligodendroglioma (T19) was clustered with two high grade glioblastoma (T20 and T26). The high grade sample glioblastoma/gliosarcoma (T18) was clustered with a gliosarcoma (T28). Two subclusters group together samples T22 and T27 and samples T23 and T25 (glioblastoma), respectively. All of these sub-clusters were combined by the clustering procedure in three clusters: T22, T27, T29, T18 and T28 were assigned to cluster 1; T23 and T25 to cluster 2 ; T19, T20, T26, T21 and T24 to cluster 3 (see Figure 4.2(a)). In samples clustering we followed the empirical approach that the number of clusters should be close to $\sqrt{n}$, with n indicating the number of items to be clustered. Observing the clusters composition, a latent structure in the tumor samples can be hypothesized. To validate this hypothesis, we used χ^2 statistical test to verify if one of the three clusters contained information on one of the four tumor types. However, no cluster was able to discriminate significantly ($p > 0.05$) between tumor classes.

4.2.5.2 SAM R package *samr* was used to perform a reduction of the dataset on the basis of the tumors grade (High/Low) and the tool was run with Δ = 0.3. SAM basically works as a generalized t-test to identify genes whose average behavior in one class is statistically significantly di□erent from the one in other(s) class(es). With these settings, SAM returned a subset of 419 (8.5%) genetic components, of which 413 (8.5%) were A□metrix ID and 6 (6.7%) miRNAs. The analysis performed on the set reduced by SAM, lead to the same sample clustering obtained without SAM pre-processing. Distance between the first three samples (T19, T21, T24), the low grade tumours, is similar to the distance found with clustering alone. Therefore, even after imposing a structure in the datasets, none of the clusters can discriminate among classes. It is worth noting that Anaplastic/Non Anaplastic tumours can be identified (with and without the use of SAM) only if the dendrogram is cut in order to isolate the subcluster T21, T24. But this approach leads to 7 and 10 clusters with and without SAM preprocessing, respectively, which, starting from 12 items defeats the purpose of clustering.

Cuto selection Cuto□ threshold for identifying genes clusters was set to 2.1, after visual inspection of a diminished variability in the clusters number around this value (Figure Figure 4.2(b)).

Sample	Tumor Type	C.1 (low/high)	C.2 (anapl/non-anapl)	C.3 (gliob/non-gliob)	(glios/non-glios)
T19	Oligodendroglioma	l	n	ngb	ngs
T21	Anaplastic mixed glioma	l	a	ngb	ngs
T24	Anaplastic mixed glioma	l	a	ngb	ngs
T18	Glioblastoma/Gliosarcoma	h	n	ngb/gb	ngs/gs
T20	Glioblastoma	h	n	gb	ngs
T22	Glioblastoma	h	n	gb	ngs
T23	Glioblastoma	h	n	gb	ngs
T25	Glioblastoma	h	n	gb	ngs
T26	Glioblastoma	h	n	gb	ngs
T27	Glioblastoma	h	n	gb	ngs
T28	Gliosarcoma	h	n	ngb	gs
T29	Gliosarcoma	h	n	ngb	gs

Table 4.7. Tumor Types and Categories. Schematic summary of tumors types used in the experiments, and category used for the discriminant.

This lead to the identification of 5 clusters composed respectively by 141 (2 miRNA), 164 (1 miRNA), 34 (1 miRNA), 48 (0 miRNA) and 32 (2 miRNA) elements. These clusters were analyzed to verify if they were significantly enriched in any GO functional category. No functional terms were found statistically significant.

4.3 Procedure

4.3.1 Dataset

In this work, we applied FA to the dataset from [LPP+07]. These data consist of 12 microarray samples (for mRNA genome-wide expression, around 14,500 coding genes) and 12 real-time PCR (for the profile of 93 miRNAs), performed on the same 12 human primary brain tumor biopsies (details in Table 4.7). On this test case dataset, we first identified the best FA model (i.e. the appropriate number of factors) based on the model's ability to explain the relevant clinical and histopathological information.

Next, we characterized the factors based on 3 properties: 1) their ability to discriminate among tumor types -this was done using Linear Discriminant Analysis (LDA, [JW02]), a supervised classifier able to find the linear combination of factors which best separates two predefined classes; 2) their functional biological characterization with the help of literature and databases; 3) their complex biological characterization, by searching novel properties emerging from the joint analysis of miRNA and mRNAs. The procedure is summarized in Figure 4.3.

4.3.2 Data preprocessing

Data from [LPP+07] were transformed by computing $\log_2$ of the intensity value of mRNA expression (miRNA data come already in $\log_2$ from real-time PCR). Quality selection filtering was performed removing every row (mRNA or miRNA expression across 12 experiments) with maximum fold change below 2.5; this reduced the dataset from 7,182 IDs to 4.966 IDs. The filtering was decided to select genetic elements with strong signal of variation. This criterion was selected as natural consequence of the filtering performed by the authors of the dataset [LPP+07] that used the same conditions to reduce the number of the IDs. Data were also normalized in different ways according to:

- $\hat{x}^1_{ij} = \frac{x_{ij} - m_i}{M_i - m_i}$, were M_i and m_i are the maximum and minimum values in the ith row, and x_{ij} is the expression of gene i on sample j.
- $\hat{x}^2_{ij} = \frac{x_{ij} - m_i}{M_i - m_i} + \mu_i$, where μ_i, is the average expression level in the ith row, and x_{ij} is the expression of gene i on sample j.

The two methods map the expression level in an interval comprised between 0 and 1 the first and μ_i and μ_i + 1 the second (in order to introduce in the model also the difference in

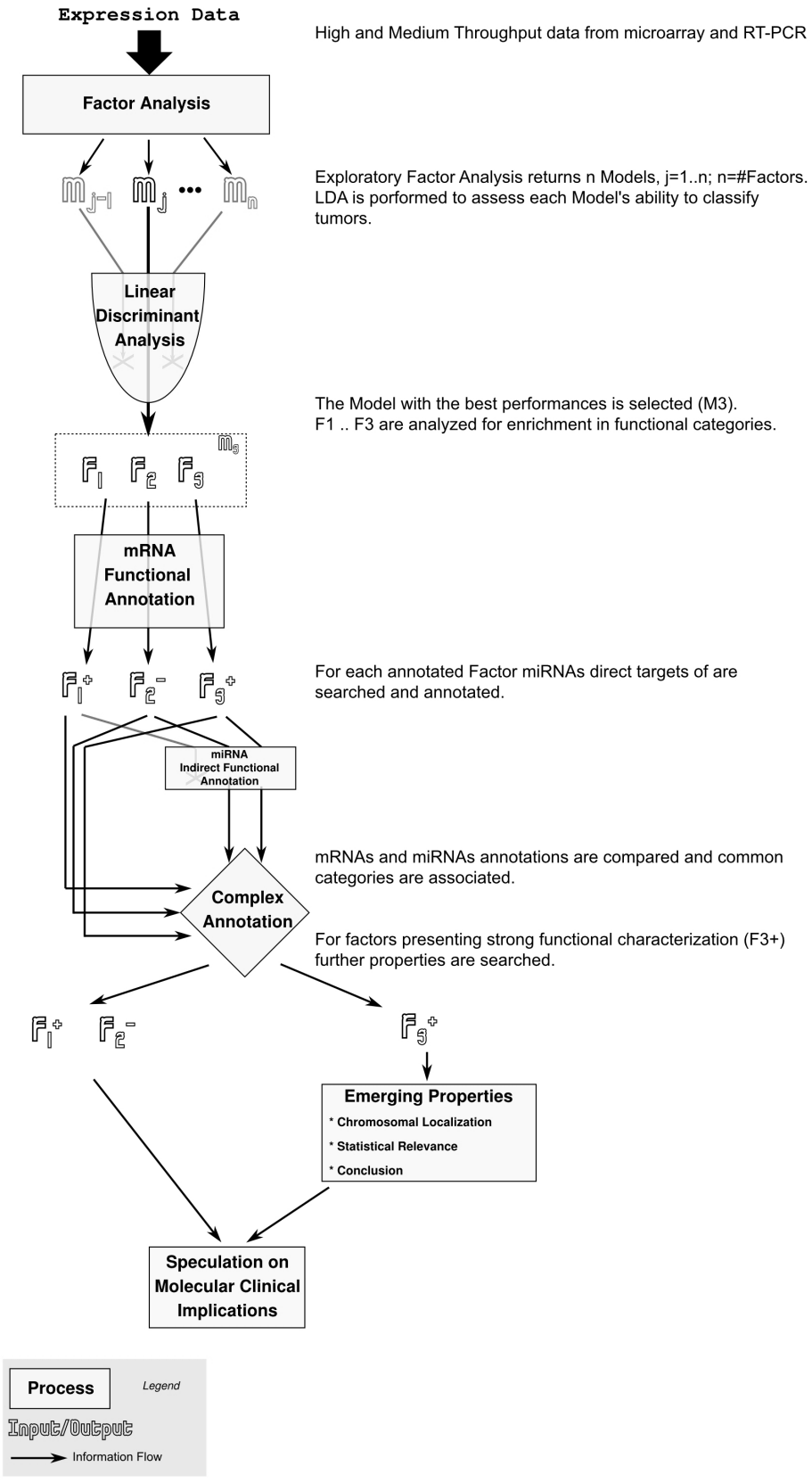


Fig. 4.3. Schematic view of the complex analysis performed jointly on mRNA and miRNAs from the same 12 tumor samples.

expression between genes). The two normalizations give identical results in the Factor Analysis step as expected. In fact, expression signals obtained from qPCR are different from signals obtained from microarrays due to the extended dynamic range of the former. It is common [WNK+07, WAC07], in order to validate a set of coding genes obtained by microarray, to express the mRNA level in each sample as a fraction of the expression level in the sample in which that mRNA is most abundant. So, from this point on, miRNA and mRNA expression data were analyzed together, as a single expression table with normalization $\hat{x}^1_{ij}$.

4.3.3 Factor analysis

The Factor Analysis model is defined in matrix notation as: $D = LF + \varepsilon$, where $D(m \times n)$ represents the data matrix, $L(m \times l)$ is the factors *loadings* matrix, $F(l \times n)$ is the factors *scores* matrix and $\varepsilon(m \times n)$ is the *unique factors* matrix. Furthermore, m are the number of samples, n the number of genetic elements and l the number of factors. Our model assumes that F and ε are independent, $E(F) = 0$, and $Cov(F) = I$. Under these conditions $Cov(D) = LL^T + Cov(\varepsilon)$, for the sake of clarity LL^T is named *communality* and $Cov(\varepsilon)$ *uniqueness*. Variability in a human tumor expression dataset arises from several sources besides tumor type, including human variability (sex, age, race) and experimental variability (systematic and stochastic errors). Available information is about tumor types, therefore, our model explicitly involves tumor types variability, and groups other causes within the ε term, showing the power of the FA method. In our work, we were interested in discovering the hidden or latent structure within tumor types, therefore FA is applied using the model $D = X^T$.

The R-package HDMD developed by Lisa McFerrin at North Carolina State University was used to take advantage of the principal axes algorithm. Communalities were estimated by iteratively updating the diagonal of the correlation matrix and solving the eigenvector decomposition. Axes were rotated to simple structure using the PROMAX algorithm to improve their interpretability. The simple structure obtained after rotation meets the requirements proposed by Thurstone [THU46, THU47] to assure the stability of FA results.

The factor score matrix was analyzed for each of the 5 models (from 1 to 5 embedded factors). The scores associated to the genes within each factor were ranked in descending order. All 3 factors presented a similar scores distribution with average $\mu \simeq 0$ and standard deviation $\sigma \simeq 0.75$. Selection has been performed by looking at the value distribution of each row of matrix F and then considering as genes associated with a factor only those whose corresponding score is outside the 2σ interval. In this way, only genes with a strong relation in the same factor were selected.

4.3.4 Discriminant analysis

The factor *loadings* coefficients matrix of each model was used to perform LDA. Four dichotomous categories (given by a class and its negate, e.g. glioblastoma/non-glioblastoma etc.) were defined (Table 4.7). LDA was also performed to assess the most likely class of sample T18 which had an ambiguous classification (glioblastoma/gliosarcoma), see Table 4.8.R-package *MASS* [VR02], function *lda()* configured to perform a classical cross-validation classification (jack-knife method, also known as *leave-one-out* validation) was used. In particular we used a *step-wise greedy* strategy, i.e. checking performances with one factor, and adding another factor, iteratively. All possible equivalent combination of factors were tested, and the most performant with the smallest number of factors involved was chosen.

Particular care must be taken with respect to sample T18, which comes in [LPP+07] with an ambiguous classification of *glioblastoma/gliosarcoma*. Gliosarcomas are variants of glioblastomas where the sarcomatous element is mixed with the gliomatous. For this reason we defined T18 as gliosarcoma, as this appears to add specificity to the glioblastoma histopathological classification of the sample. However, to disambiguate its definition, and to assess the importance of this classification in the final results on Model 3, we assigned T18 to both categories

(a) Tumor Grade

	High/Low Grade	
	P High	**P Low**
High	**9**	**0**
Low	**1**	**2**
	p=0.021	

(b) Anaplastic

	Anaplastic/*Anaplastic	
	P Anap	**P *Anap**
Anap	**10**	**0**
* Anap	**0**	**2**
	p=0.004	

(c) Glioblastoma

	Glioblastoma			
	P Glio	P *Glio	**P Glio**	**P *Glio**
Glio	7	0	**6**	**0**
* Glio	2	3	**1**	**5**
	T18 as Glio		**T18 as *Glio**	
	p = 0.072		**p=0.012**	

(d) Gliosarcoma

	Gliosarcoma			
	P Gsar	**P * Gsar**	P Gsar	P * Gsar
Gsar	**3**	**0**	0	2
* Gsar	**1**	**8**	0	10
	T18 as Gsar		T18 as * Gsar	
	p=0.02		p = 1	

Table 4.8. Discriminant Analysis on Sample T18. These tables show the classification performances of Model 3 when T18 is classified in 4 different possible ways: Gliosarcoma and dual, and Glioblastoma and dual. Significant classifications ($p < 0.05$ after Bonferroni correction) are shown in bold. Classification performances are statistically significant when T18 is classified as Gliosarcoma and not as Glioblastoma. Anap: Anaplastic; *Anap: non Anaplastic, Glio: glioblastoma; *Glio: not glioblastoma, Gsar: gliosarcoma; *Gsar: not gliosarcoma.

and their complement (Gliosarcoma/Non Gliosarcoma; Glioblastoma/Non Glioblastoma) and then performed the LDA on the 4 di□erent datasets obtained. If LDA is able to correctly discriminate the known classes, we infer it can also properly assign this ambiguous sample.

As shown in Table 4.8c) and Table 4.8d), LDA performs a better discrimination when T18 is assigned to the Gliosarcoma class rather than to Glioblastoma. When the sample was assigned to the category of the Non Gliosarcoma, LDA lost completely the capacity to detect gliosarcomas in the samples. Conversely, when T18 was assigned to the Glioblastoma class, LDA assigned 9 samples instead of 7 to Glioblastoma.

4.3.4.1 Alternative Classification Analysis Loading scores contain a signal that a classifier can extract to discriminate between tumor classes. We used 4 standard alternative classifiers: Support Vector Machines, Neural Networks, Naïve Bayes and K-Nearest Neighbor downloaded from `http://www.patternrecognition.co.za/sourcecode.html`. We used the default parameters to classify the four tumor classes.

4.3.5 Model selection

To evaluate the performances of each factor model on the four tumor classes, we evaluated the contingency table obtained from the discriminant analysis by Fisher's exact test. The null hypothesis assuming that the discrimination between two tumor classes is due to chance was rejected for $p < 0.05$. For models with similar prediction scores we kept the one with fewer factors.

4.3.6 Functional classification

On both FA and clustering (used as alternative method to our approach, see Supplementary Information) functional analysis was performed using the online tool DAVID [DSH+03, HSL09] using GO terms, Kegg pathways terms, SP keywords and features and InterPro terms. The whole list of 4876 probe ID was used as background population. In order to reduce the number of non significant associations, a resulting functional cluster was further analyzed if and only if it contained at least one category with Benjamin score < 0.05. The indirect functional analysis performed to describe miRNAs relevance was performed by searching manually in TarBase all the known coding genes that are target of the miRNAs identified by the FA and clustering.

Then for each gene a list with all the associated GO terms was compiled. Due to the small number of targets obtained no p-value could be associated to any GO term.

4.4 Conclusion

The capability to discriminate between *a priori* defined classes can be achieved in a variety of ways (a comparison with supervised and unsupervised algorithms is provided in the Supplementary Information). However, the capacity to generate factors explaining the complexity of the molecular interactions requires the ability to construct multilevel clusters. With the data at hand we showed that this cannot be achieved in parallel analysis (versus simultaneous or joint) of the two datasets (mRNA and miRNA) or with other approaches we evaluated. The interpretation of factors based on associating them to mRNA/miRNAs represents the major contribution of this work. Certainly, the study of [LPP+07] shows sample size limitations (12 patients enrolled) therefore our analyses must be considered as an exemplar of the factor analysis approach.

Globally, based on this analysis, since the miRNAs in $F3^+$ belong to two redundant *clusters* of miRNA, we can speculate that: 1) one of the biological functions in which these *clusters* could be involved is the regulation of the transcription and 2) in some way, in brain tumors these two *clusters* are active whereas, in normal cells, only miR-17-92 appears to be constitutively expressed. Probably both *clusters* act on the same set of coding genes, but the two loci are regulated separately in normal cells [LLLR07]. Nevertheless, despite this strong relationship between the 2 *clusters* it is difficult to understand how this redundancy works effectively in cells. However, the finding of a possible activation of the polycistronic genes miR-17-92 and miR-106-363 represents an encouraging evidence that the factorization of the miRNA and mRNA data can reveal latent structure in the configuration of the expression levels in tumor samples.

Despite obvious limitations, we believe our results clearly show that this approach is a very powerful one for the study of multilevel *omic* data, which in turn can bring more insight into understanding the complex mechanisms of the transmission of information in the cell as a whole.

Chapter 5

Insertion Sites Detection

5.1 Motivation

Gene therapy is a therapeutic approach based on the principle that a defective gene can be corrected using a healthy gene copy stably introduced into the host. Retroviral vectors which stably integrate as proviruses into the cellular genome have been successfully applied for the correction of hematological diseases [CCLHBAF05, Koh08].

From insertion site analyzes of preclinical studies and clinical monitoring of gene modified hematopoietic stem and progenitor cells in patients it has become evident that biologically relevant vector induced side effects ranging from in vitro immortalization to clonal dominance and oncogenesis in vivo accompany therapeutic efficiency of integrating retroviral gene transfer systems. Analyzing the insertional repertoire of treated patients and newly established therapeutical vectors in preclinical animal models has become an important parameter for dissecting the genotoxic potential of retroviral vectors.

A specific list of risks and limitations could be associated to gene therapy procedures [BKM+06]. The a-priori probability that an insertion event disrupt a tumor suppressor genes or trans-activate a proto-oncogene has been considered negligible. Initial reports from the SCID-X1 gene therapy found that a Murine leukemia Virus (MLV) derived vector was integrated in proximity of LMO2 proto-oncogene in 2 out of 10 treated patients.

It has been described that in addition to overt carcinogenesis, even as a single vector copy, many insertion locations more subtly influence the biological fate of a cell clone in vivo. With LAM-PCR technology [SCS+03, SSB+07] specifically developed for this purpose, it has been shown that over 40% of circulating cells carry insertions in a rather small set of frequently affected common insertion sites (CIS). Such CIS are almost always in the direct vicinity of known and novel genes likely involved in cellular growth, survival and self-renewal processes of immature progenitor and stem cells [HBAVKS+03a, HBAvKS+03b, DHBAS+07, SHS+07]. Insights into the distribution of retroviral vector integration sites in human gene therapy patients suggest that most of the insertion loci which are of biological relevance are not located in unchartered genetic territory but harbor a “known suspect” gene waiting to be discovered in the act of distorting cellular survival [HBAVKS+03a, HBAvKS+03b, SCS+03, DHBAS+07, SHS+07, BSS+10].

First analysis of integration sites have been performed using inverse-PCR, LM or LAM-PCR combined with shotgun cloning and Sanger sequencing [DHBAS+07, ACA+07]. This approach is expensive, time consuming and low sensitivity with only 100-300 insertion sites per patients.

To overcome these limitations Schmidt et. al [SSB+09] recently introduced a new technique that, combines LAM-PCR with next generation high-throughput sequencing platforms allowing a comprehensive assessment of vector insertion sites and semi quantitative estimations of clonal dynamics and hematopoieses by measuring the relative retrieval frequency of individually retrovirally marked clones following sequencing [GEP+09, PAG+10]. In combination with optimal restriction motif usage as well as the establishment of a novel non-restrictive LAM-PCR approach the genomic accessibility to vector insertion sites has further been improved substantially however sampling error as well as amplification biases can still influence the ac-

cessibility and quantification of insertion sites making sensitive capturing techniques aimed to overcome amplification bias highly attractive for clinical monitoring of gen therapy patients.

In this project we aim to establish and optimize a novel high-throughput sequencing strategy for retrieval of insertion sites which is independent of PCR amplification. The capturing of integrated vector insertion sites using sate of the art array based hybridization steps was evaluated for further optimization. Hybridization based methods allowing a direct selection and next generation sequencing of whole integrated vector copies enable the precise analysis of vector insertion sites as well as true germline configuration and stability of therapeutic vectors in clinical and preclinical studies. The direct selection of virtually any type of sub genomic elements followed by next generation sequencing of the genetic information is of high value for genome and cancer genome characterization. Once proven successful for precisely identifying vector insertion sites including simultaneous identification of both flanking genomic regions and whole interior vector sequences such technical platforms can be exploited for analyzing chromatin marks of integrated vectors and surrounding cellular sequences in combination with chromatin IP procedures. For evaluating the feasibility of this approach human cell lines, which were transduced with lentiviral vectors and in which insertion sites have been characterized by nr-LAM-PCR and 454 sequencing previously, were subjected to direct selection techniques based on hybridization (FEBIT) and next generation sequencing (SOLiD).

The aim of project was to establish bioinformatical platforms required to filter and resolve raw sequencing data obtained from the sequencing of captured vector genomes and insertion sites by the hybcapture method o□ered by the FEBIT company using specially customized chips.

5.2 Procedure

5.2.1 Genomic data

Geniom platform (FEBIT; `http://www.febit.com/`) provide a highly automated enrichment environment for DNA fragments. The DKFZ team and the FEBIT company developed a biochip in which each spot is a target for a specific vector's region. Their customized chips were developed completely covering the vector by 1 bp spaced 50 mer probes. For maximizing retrieval of flanking cellular sequences, additional 50 mer probes for the LTR sequences were added so that close to 47% of the surface spots is covered by 5' and 3' vector's end (Figure 5.2).

Two micrograms of DNA was extracted from a sample of human fibroblasts transduced by lentiviral vector. The DNA was captured and sequenced ocurding to FEBIT.

5.2.2 Sequencing and filtering procedure

As specified in Sasson et al. [SM10] SOLiD reads are a□ected by both polyclonal and independent errors.

A set of 43,487,758 SOLiD sequencing reads of 50 base length were analyzed accordingly with a filtering procedure based on the values provided by the SOLiD quality file that reduce noise in data and computational resources. We used `SOLiD_preprocess_filter.pl` a PERL script, provided as supplementary material by Sasson ([SM10], `http://hts.rutgers.edu/filter/`), modifying the input parameters (see Table 5.1) in order to reduce the stringency of the filtering procedure and maintaining the mean quality in all positions reliable (quality value > 20; Figure 5.1).

5.2.3 Align in color space

Based on design and previous characterization of the sample analyzed some expectation on general results were anticipated. These include the specific retrieval of defined genomic regions containg known insertion sites in the EPB41L2 and CCR5 gene (personal communication). In a

```
SOLiD_preprocess_filter.pl -f sample.csfasta -g sample.qual -q 20 -e 6 -o outputFile
```

Table 5.1. Parameters used to filter the SOLiD reads. -f: csfasta file for the forward strand; -g: quality file for the forward strand; -q: polyclonal analysis minimum QV score (default 25); -e: the maximum number of errors allowed per read (default 3); -o: prefix of the output files name.

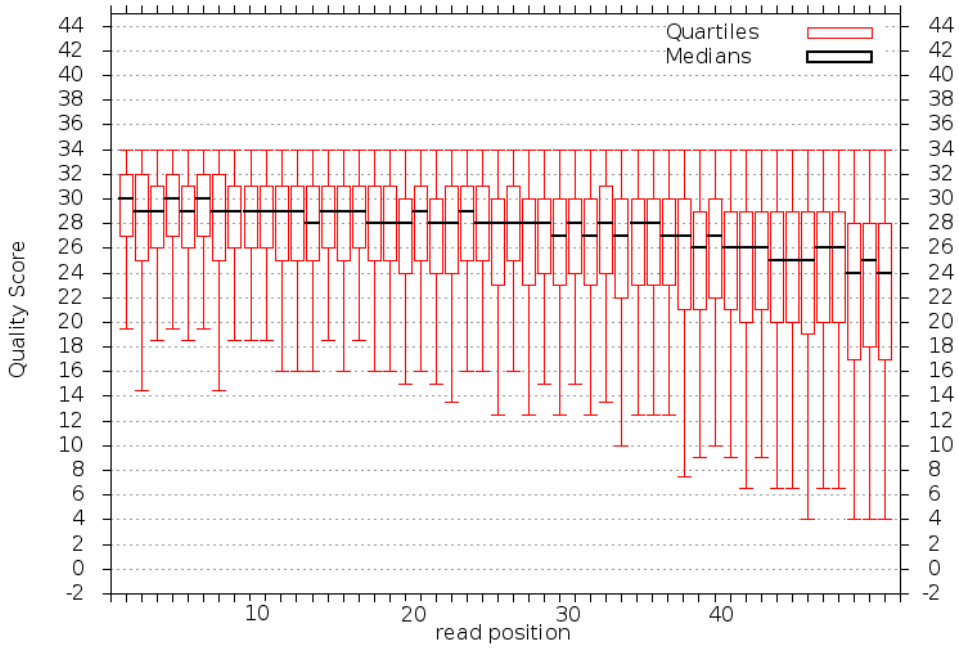


Fig. 5.1. SOLiD quality plot after filtering procedure. As expected the quality decrease in the terminal portion of the reads but maintaining a good median quality (24 in position 48 and 50).

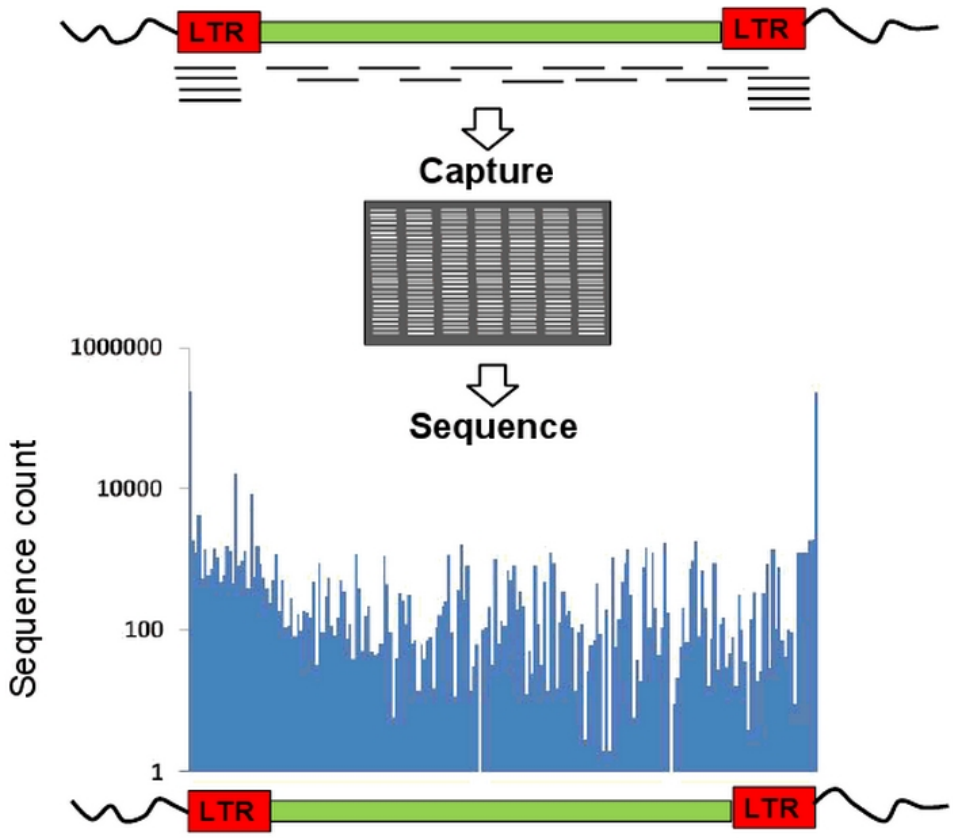


Fig. 5.2. FEBIT capturing method. The whole vector is represented on the biochip with a preference for the LTR external portions. This design bias is expected to be highlighted counting the number of reads along the vector.

first step the reliable set of reads was aligned to the vector expecting that most of the coverage was detected on the boundary.

For this step MOSAIK was used (see Table 2.3) which contains an option that permits to align SOLiD reads in color space via previous conversion of the target in color code. Is important to note that the two most important regions for defining insertion sites based on identification of clear vector-cellular boundaries, the so called Long-terminal-repeats of the vector (LTR, see Figure 1.3), are identical in sequence masking the mapping of the vector fragment to the 5´- or 3´-LTR. This limitation was overcome by introducing a string of neutral "N" symbols from the base 100 to the base 234 and from the base 3,711 to the base 3,844.

As mentioned MOSAIKALIGNER was used for both vector and genome alignment. Program parameters are selected on the basis of the considerations in 5.2.3.1[1].

5.2.3.1 On the reads length Here we faced for the first time with the main challenge that a□ect the design of the project: the length of the reads. Usually, one of the steps involved in high-throughput insertion sites detection is the trimming of vector and link sequences from the hybrid[2] reads in order to obtain only the genomic fragment that is used for genome mapping

[1] parameters: : `-hs 5 -mm 0 -mhp 100 -act 6 -mmal -minp 0.3`
-hs: hash size
-mm: maximum mismatch threshold
-mhp: hash position threshold
-act: alignment candidate threshold
-mmal minp: minimum percent alignment threshold

[2] When a read contain both genome and vector sequence we call this read *hybrid*.

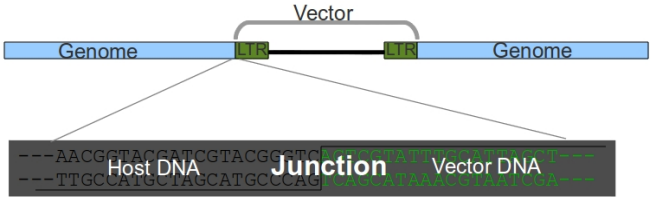


Fig. 5.3. Vector-Cellular junction. The vector portion is removed from the read and the flanking region is aligned and annotated on the target genome.

and thus retrieve information on the targeted loci (Figure 5.3). Because the kinetic complexity of the vector is 4,000 bp[3] we could imagine that is possible to map uniquely the vector using only sequences with length 6([4]). We used this characteristic to set the aligner's parameters. The complexity of the human DNA, used to query the Geniom Biochip, is about $3 \cdot 10^8$ taking into consideration only the slow kinetic component. We expect by chance, under these assumption and under a simple uniform model, more than 1 match if the query sequence is 14 bases-long. Considering that the vector "generate" 4,000 di□erent strings we need a match composed by at least 20 bases[5] to have a probability equal to $1/4000 \approx 0.00025$ to retrieve random match when the kinetic complexity is comparable with our data. Therefore if the sequence has a total length equal to 50 both the genomic and the vector ends must be composed at least by 20 bases.

5.2.3.2 Vector mutation The Geniom Biochip is designed to contain about 1Mb of target region and guarantee high coverage (`http://www.febit.com/microarray-sequencing/`). The SOLiD color code in combination with high coverage is useful for SNP calling. We exploited this feature to: 1) check the global technical quality of the experiment and 2) detect eventual mutations appeared on the vector.

The captured target region is relatively small (about 4000 kb) expecting us to obtain a very high coverage. The organization of the vector on the chip is asymmetrical because biologists decided to use the 47% of the synthesis space for 500 bases that belong to the two ends of the vector and the remaining 53% for the other 3450 bases. The awaited coverage ratio between the two chip areas is computed as $R = \frac{0.45}{0.55}\frac{3450}{500} \approx 6.25$.

Browsing literature and the FEBIT web site, we found a published paper [SWH+09] and some reports where the mean coverage and the target region length for four experiment are described (Table 5.2). From this data we could compute the expected mean coverage on our vector. Giving a first target region with known dimensionT_1 and coverage C_1 and a second target region with known dimension T_2, the unknown coverage C_2 can be calculated using the equation $C_2 = C_1\frac{T_1}{T_2}$. The upper bound for the coverage in our data is 125,000 on average, assuming that all the selected reads align with the vector (Table 5.2).

Mutations in the wild type retroviral genome are readily introduced and my occur at di□erent sages of the viral replication cycle. Retroviral based vectors are defective in replication and only produce a single round of integration, however mutations may still be introduced during reverse transcription and cellular replication e□ecting stability of vector genomes. Due to the fidelity of cellular DNA replication (10^{-9} to 10^{-11} substitution per base pair), it is unlikely that mutations occurring during cellular replication of the provirus contribute significantly to

[3] An approximation based on that the vector contains only unique sequences gives a complexity that is equal to the length ($\approx 4Kb$).

[4] $4000 \approx 2^{12} = 4^6$.

[5] 20 bases generate a sequence space of 2^{40} sequences. The number of expected matches when the alignment is 20 bp, the target is 4000 bp and the complexity of the query is $3 \cdot 10^8$ bp is

$$\frac{3 \cdot 10^8\ 4 \cdot 10^3}{2^{40}} \approx 1$$

Case study	*Coverage (C_1)*	*Target (T_1)*	*Expected Coverage ($C2$)$^{++}$*
1*	402	498 Kb	50,049
2^{+}	176	482 Kb	21,208
3^{+}	2388	160 Kb	95,520
4^{+}	1336	160 Kb	53,440

Table 5.2. Reported coverage for some experimental results. From [SWH+09] ; $^{+}$ from `http://www.febit.com/`;$^{++}$ considering the target dimension equal to 4000 Kbases.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
XYZ	1001	A	A	11	0	15	3	..c	II
XYZ	1002	C	C	75	0	25	6	...,.,	III2
XYZ	1003	T	A	255	238	255	6	a..aa.,	III7
XYZ	1004	G	G	255	0	255	5	.....	III1

Table 5.3. 10 columns MAQ model explanation. 1 Reference name; 2 position; 3 reference base at that position; 4 consensus bases; 5 consensus quality; 6 mutation quality; 7 maximum mapping quality; 8 coverage (# reads aligning over that position); 9 bases within reads where; 10 quality values (phred33 scale, see Galaxy wiki for more).

the high degree of vector diversity. Therefore, most mutations are introduced during reverse transcription and/or RNA pol II transcription of the provirus. In our experiment the last possibility is not a source of variation due to the defective nature of the vectors used in gene therapy

The analysis of the coverage on the vector sequence could reveal (with the coverage expected in Table 5.2) mutations with a high degree of confidence. The analysis started using the resulting alignment file converted in BAM format (using MosaikText [6]). Then SAMtools [LHW+09] is used to generate a pileup file according to Maq model [LRD08]. The pileup file is a flat file where each row represents a position on the reference sequence (the vector sequence) and the columns represent information shown in Table 5.3 on that location . Columns number 2,4,6 and 8 that represent position, variation, mutation quality and coverage respectively, are used to filter out mutations. The deep of the coverage is really high and a considerable fraction of positions contain significant mutations with MAF < 0.01. For this reason we decided to increase the stringency and to report only mutations where the MAF is more than 0.01.

5.2.3.3 Align on Genome The human genome was converted in color space using MosaikBuild in order to map the hybrid reads obtained aligning the vector. Then the alignment was converted in a tabular SAM format and checked for the Hypothetical insertion sites found on a local version of the Ensembl Genes version 59 (`http://www.biomart.org/biomart/martview`).

5.2.4 Align in base space

Although contiguous errors will arise from a single color sequencing error when the sequence is decoded before the alignment, we decided to remove the constrain imposed from the coding nature of the SOLiD reads and to translate all the reads in base code (Figure 5.4).

5.2.4.1 Capture hybrid sequences Then we used a simple script that search, via regular expression, the first l bases of the vector (the pattern; Figure 5.5) . Vector bases are then trimmed and the remaining portion of the read is a candidate to detect the target genomic region. The dimension of the pattern search window was selected on the basis of the considerations in 5.2.3.1 and after an heuristic exploration of the l parameter space. This exploration

[6] parameters: `-in alignmentName -bam outBamName`

```
BEGIN{
                flag=0;
                while(getline < "klein.txt" > 0) {
                        lead=$1;
                        # Load the matrix and assign the
                        # the corresponding code number
                        for (i=1;i<5;i++) {
                                mat[lead,i]=$i
                        }
                }
        }
        {       if (match($0,"#")) { # Remove comments
                        flag=0
                }
                else if (flag==0) { # the line is the header
                        flag=1;
                        print $0
                } else {        # the line is the sequence
                        flag=0;
                        split($0,color,"");
                        val=color[1];
                        printf("%c",val);
                        # reverse the sequence
                        for(i=2;i<=length($1);i++){
                                val=mat[val,color[i]+1];
                                printf("%c",val)
                        }
                        print ""
                }
        }
```

Fig. 5.4. awk conversion script between SOLiD color code and base space. The conversion matrix `klein.txt` has this structure $\begin{bmatrix} A & C & G & T \\ C & A & T & G \\ G & T & A & C \\ T & G & C & A \end{bmatrix}$. The script uses the the color code at the position i and $i-1$ as hash code to address the base in the matrix. It translate about 87 Mb per minute per CPU.

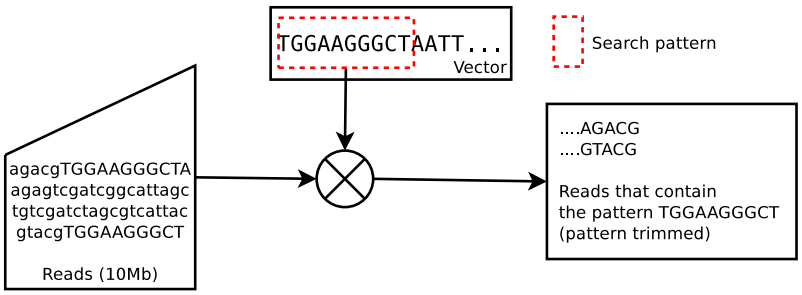


Fig. 5.5. Diagram for reads selection in base space. First the pattern is searched and then trimmed.

A

```
11     1  1     1       6      1      1  2
AACGGTACGATCGTACGGGTCAGTCGTATTTGCATTAGCT

  2      2  1  1 1       2  3         5
GTAGCCCTAGCACTACACGGGTCAGCAGTATTACGAGTCG

1  1       1    1    2     1 1      1      3
ACTACACGGTACGTACGCAGCAGATACGTCGATATCGAGT

   11      1     2      1  1      11     1  1
TCTAATTTGTATCTCAGCCTGATCATGTGTAATTCCATCA
```

B

Pattern Length	Pattern	# of Match Observed (O)	# of Match Expected v	# of Match Expected r	Ratio (r/O)	Ratio (r/v)
1	A	36	43	40	1.11	0.93
2	AG	10	12.93	10	1	0.77
3	AGT	4	5.35	2.5	0.625	0.47
4	AGTC	2	3.4	0.625	0.3125	0.18
5	AGTCG	2	2.86	0.15625	0.0781	0.05
6	AGTCGT	1	2.66	0.0390625	0.0391	0.015

Fig. 5.6. Effect of pattern length l on the number of matches. D is a sample of m-long reads, that contains vector sequence starting from position i (where $i = 1..m$) and R is a sample that do not contain vector sequences. Under the uniform random model, $r = mN(D+R)\cdot 4^{-l}$ is the number of sequences extracted by chance using the first l characters of the vector, where $N(\bullet)$ return the number of reads contained in a dataset. If the distribution of the starting points of the vector is uniform in D we could assume that the number of reads extracted from D is $v_n = (1 - \frac{l-1}{m})N(D)$. Using these relation we can compute the expected number of observations in a mixture model using in this equation $v = r + v_n = mN(D+R)4^{-l} + N(D)(1 - \frac{l-1}{m})$. Is also possible to compute the expected ratio r/v using the rearranged equation $\frac{r}{v} = \frac{1}{1+\frac{N(D)(m-(l-1))}{m^2 N(D+R)4^{-l}}}$. In A three sequences are hybrids sequences (vector in red) and one is random.The length of each read is 40. In B are listed all the possible patterns from the vector with a window that vary from 1 to 6 bases. As is shown, despite the small sample, the observed r/O and the expected ratio r/v follow the same behavior. The numbers over the sequences flag which is the pattern that match on that position.

provides to count the number of the reads returned from the "capturing" procedure varying l from 4 to 20 with step 2. The idea is based on the assumption that the increment of the number of bases in the pattern return a number of reads that decrease 1) exponentially when the pattern is too short and 2) linearly when the pattern length is su□cient to detect non random hybrids (see Figure 5.6 for explanation). With this model in mind we plotted the number of reads returned versus the corresponding l; then a minimum width was selected manually where the behavior of the graph translate from exponential to linear (l = 10 in our case, Figure 5.7) .

5.2.4.2 Align on genome The set of reads that contains (trimmed) terminal portion of the vector are aligned on the human genome (version hg19) using MEGABLAST [AMS+97] that is fast, reliable and do not require quality values to align sequences. The selection of the alignment parameters was a compromise between the need of insertions site and the reduction of noise. As discussed in 5.2.3.1, at least 20 bases are required to target a unique site in the human genome for most of the alignments. A word size with length equal to 16 guarantee that pieces of putative genome have a seed sequence that has a perfect match with al least 16 bases of the query (the read). Then the expected value was set to $1\,10^{-7}$ and the identity score to 90%. Whit this setup MEGABLAST discard reads that are to small or with an excess of mismatch and return a set of calls that are good candidates to be real insertion sites.

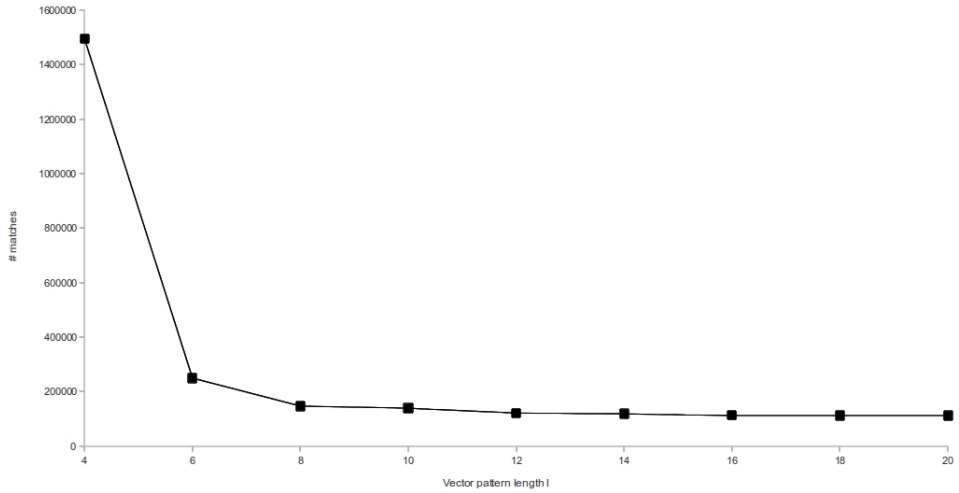


Fig. 5.7. Pattern length selection graph. When $l > 10$ the the exponential behavior (random selection) is negligible.

Before the "functional" annotation we filtered out reads that map in more than one location on genome with the same score using a python script developed for this purpose.

5.3 Discussion

From the starting population composed by 43,487,758 reads 10,911,004 of them (25%) passed the filtering procedure. A previous analysis, with the filtering procedure and default parameters, returned less than 6 million of reads (1/7 of the entire set) . For this reason we decided to modify the default parameters due to their excessive stringency. We think that our setting is a good compromise to obtain a data set with high Signal to Noise ratio but where sensitivity is not compromised.

5.3.1 Vector alignment

The alignment procedure on the vector was fast, due to this small length, allowing fine tuning of the parameters and minimizing the number of failed hash code. Theoretically with 6 bases all the possible configuration of a vector are correctly sampled; due to unavoidable sequence redundancies in the vector we found that using a hash code composed by 5 elements the number of unique reads mapped reached a maximum of 3,188,420.

At this point we verified: 1) if the coverage obtained was in the same order of magnitude of the expected coverage in Table 5.2 (the expectation interval goes from 21,208 to 95,520); 2) if the distribution of the coverage on the vector is similar to the coverage which should results from the chip design. The coverage distribution is reported in Figure 5.8.

The observed average coverage (26,189) fall in the low portion of the interval and shows that our raw results are in agreement with previous findings although a bit less. One explanation for this attenuation is the filtering procedure that discard 1/4 of the reads and reduce the coverage; but also the capture method or other factors could be responsible for this phenomenon.

The coverage distribution ratio R between the LRT and the internal vector regions is 4.3 whereas the expected one is 6.25. This ratio has a value that is a little bit lower than expected. The first explanation is that the area of the chip covered by the vector ends is 38% (instead of 47%) and the second explanation is that the e□ciency of the capturing method was not

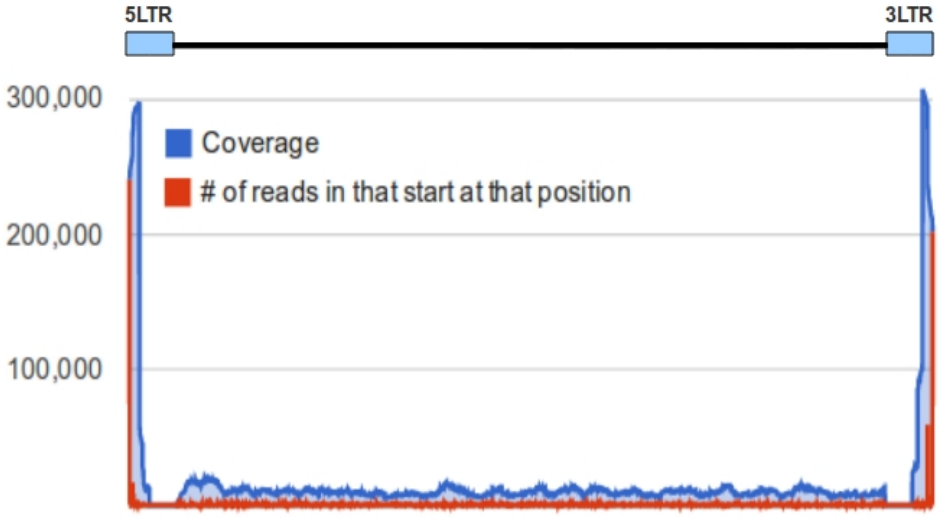


Fig. 5.8. Coverage and Start Site distribution on vector. The two holes at the beginning of the 5LTR and 3LTR are a artifact due to the introduction of "*N*" symbols in the vector sequence to avoid conflict during the alignment.

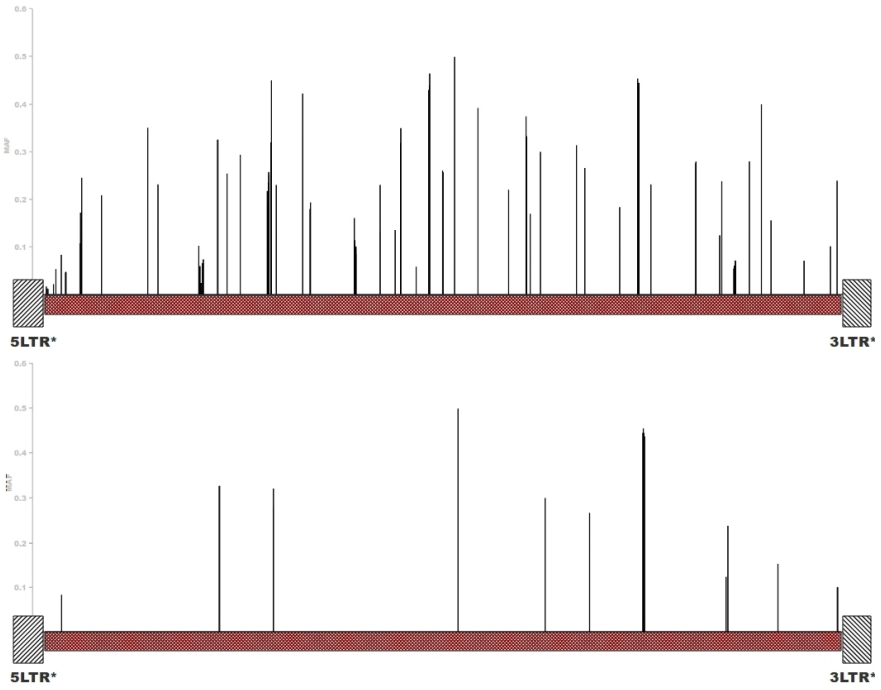


Fig. 5.9. Raw and filtered vector mutations. Black 45 degree lines: 5'LTR from position 1 to 100; Black -45 degree lines: 3'UTR from position 3846 to 3945; red crossed 45 degree: internal sequence from position 235 to 3710. * position from 101 to 234 and 3711 to 3845 (with 'N' letters) are removed from the picture. In the LTRs no mutations are detected.

uniform on the vector. We lean for the second explanation because the first means improbable errors in chip design and/or probes synthesis.

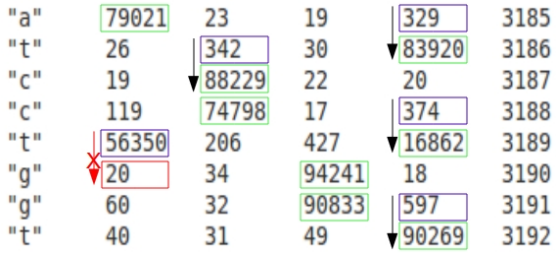


Fig. 5.10. Mutation matrix example. Green rectangle: "master" base; blue rectangle: "slave" base; red rectangle: wrong expected "master" base at position $p+1$; arrows: expectation step between position p and $p+1$.

5.3.2 Mutation detection

In total we found 99 point mutation that occur all over the vector length except in the LTR regions with a mutation frequency that correspond to one mutation every 40 bases. As is shown in Figure 5.9 the ratio between the consensus and the target base (MAF) increase along the first part of the vector; most of the mutations cluster together and only 1/3 of them are not in cluster. The absence of mutations in LTR is explainable on the fact that mutated LTR are most likely not functional in terms of the integration reaction catalyzed by the viral Integrase gene, producing vectors that are defective for the integration. Cluster of mutations are a phenomenon which hide a more profound significance. One explanation could be that some positions are preferred mutation hot spots in the vector but this needs to be analyzed more deeply by comparing mutation hot-spots in the vector in biological replicates.

A second explanation lie on how the SOLiD sequences are aligned on the vector. To understand our point we need to introduce a behavior observed in the mutation matrix that contains, for each position of the vector, the number of reads that have an "A" a "C" a "G" and a "T". We realized that is possible to predict the next base in the sequence simply observing the bases frequency. Giving the position p the base with the bigger frequency, the "master" base, is the actual base and the second biggest one, the "slave" base, will be the consensus base in p + 1. This behavior is a source of noise because when the frequency of the slave base is high or a region is covered with high quality reads, the mutation caller program detect one or more mutations where no mutation is present (Figure 5.10).

We exploit this comportment developing a program that predict the base p + 1 and flag those position where the expected base in p + 1 di□er from the observed one. This script and the mutation caller combined together discard all those mutation that are predictable on the basis of our model.

In Table 5.4 are listed 16 vector variations that come from the combined filtering strategies. A deeper analysis in the significance of these mutations will be necessary in order 1) to understand the source of this deterministic behavior and 2) to see if some of them could modify the vector functionally.

5.3.3 Genome alignment

5.3.3.1 In color space All the 3,188,420 reads that map on the vector were parsed to keep the hybrids that start on the first or on the last position of the vector (438,504; 13.8%). The mapping on the human genome (hg19) was done with MosaikAligner[7] aligning 7,496 (1.7%) reads in 4 minutes. Each mapped position was joined with the human genes positions in order to detect potential genes that are involved in the integration process. Most of the reads correspond to intergenic regions and only 66 insertion sites fall in gene regions (36 genes), giving less than 2 insertion sites per gene on average. This result was not satisfactory for two main reasons: 1)

[7] parameters: `-hs 15 -mm 2 -mhp 100 -act 20 -mmal -minp 0.3` (see Note 1)

Position	*Vector Base*	*Consensus Base*
253	c	t
949	c	g
1186	t	a
1187	a	g
2001	g	c
2383	t	g
2580	g	a
2815	t	c
2816	c	g
2818	a	g
2820	g	a
2821	a	c
3180	t	c
3189	t	a
3409	a	c
3672	c	g

Table 5.4. Vector Mutations.

in the gene list do not appear EPB41L2 and/or CCR5 and 2) the number of insertion sites detected by this methods is undersized. Certainly, with the length of our reads we expected that the sensitivity was lower than the sensitivity of the LAM-PCR but this result did not mach our quality criteria and for this reason we undertook the way that perform all the steps in base space.

5.3.3.2 In base space As described in 5.2.4.1 after the translation of 10,911,004 reads in base space more than 138,000 reads were selected (in the forward direction) and trimmed. These set was aligned against the human genome using MEGABLAST[8] and the result, 179,976 positions, was filtered in order to maintain only reads with unique mapping. The 161,541 unique reads were further filtered to keep only the alignments in which the first base involved in the match was the base at position 1. This last step provided 106,796 potential flanking regions found in the human genome.

More than 80,000 (75%) sites fall in 79 genomic regions. In Table 5.5 the genes are sorted on the calls (number of insertions sites). The first two position are occupied by EPB41L2 and CCR5 with 79,046 and 821 calls respectively.

5.4 Conclusion

The performance of the methods that manage sequences in base space worked better. This is disputable since literature and technical papers warn researchers involved in SOLiD technology to avoid to decode color reads before the alignment. Our findings are probably explained considering that SOLiD system works better in SNP detection, when one base mutation leads to two contiguous color changes, but not when one read contains two kind of sequences of which one is necessary to select the read and the second to map it on a genome. Furthermore one color code sequence could potentially align with 4 different target without mismatch (more if mismatches are allowed); so when the color code reads are small as in our case, most of the alignments are due to wrong mapping.

As a control we inspected manually a small sample of the aligned reads in color code. The idea was to check if the decoded read match the portion of the gene detected by the read. The result of this test was instructive: no one of the read contains the found gene fragment.

[8] parameters: `-a 8 -W 16 -p 90.0 -e 0.001`.

Gene Name	*Calls*
EPB41L2	79,046
CCR5	821
ZCCHC14	625
EYA4	19
CCR2	9
MREG	7
PECR	7
FER	6
SUV420H1	5
Others	77
Total	80,622

Table 5.5. Genes in which insertion sites have been found. In table are reported only genes with more than 4 calls; genes with less than 4 calls are: OCIAD1, HIPK2, IL1RAPL2, PDE11A, PGM5P1, UNC5D, WWOX, ALAS2, AOAH, BYSL, C1orf175, C20orf185, CACNA2D3, CDH13, CSRNP1, DHDH, DPP6, FAF1, FAM135B, FAM180A, FBN1, FBP1, GNG12, GOPC, GPR39, GUCY1A3, IFLTD1, ITGA6 IWS1 KLHL22, LRTM1, MACF1, MCART1, MEAF6, NPHP1, P4HA2, PDE1C, PDE4B, PDE4DIP, PFN2, PGM5P2, PHACTR2, PINX1, PLCG2 PPOX PTPN4, RAB11FIP1, ROS1, RPE, RUNX3, SAMD3, SCARF2, SGCZ, SGPP2, SLC15A5, SLIT3, SMOC2, SNTG1, SOX7, TMEM104, TNS3, TRAC, TRBC1, TRERF1, TRPV3, USH2A, VPS13B, WDR72 and XPO6.

On converse when we coded the gene sequence in color code, partial matches are found. These findings indicating that the alignment in color code of short sequences could give wrong results.

As expected by the design of the vector, integration sites in EPB41L2 were found more frequently than the others and, although CCR5 is found with a strong signal, the number of total insertion sites detected is 2 orders of magnitude less than the integration sites obtained by LAM-PCR. This means that the sensibility of the method is low and that other strategies must be chosen in order to reach performances that make direct sequencing approach attractive. While the selection strategy of the library sequence could be matter of discussion we can assert that the sequencing technology must provide long reads in standard fasta format.

Appendix A

R Bus package

In this appendix I show a side work that I contibuted to develop during my stay in Shanghai and that ended with the release of a R package (downloadable from Bioconductor `http://www.biocontuctor.org`)

The number and type of data support produced by innovative biotechnologies to screen molecular biological data expands very rapidly. Indeed, it is currently possible to observe the activity (over- under- expression, presence or absence of mutations) of virtually all the molecules of a given type (mRNA, miRNA, DNA) in one single screen, through high-density chips , or via sequencing related techniques. Due to the di□erent biogenesis of these molecules and to the di□erent technologies behind their screening, preliminary analyses require customized procedures. However, once the data are processed, their structure can be reduced to a large tabular format, which rows represent molecules' activity (over- or under- expression for mRNA and miRNA data, presence or absence of mutations for DNA data), and columns represent experimental conditions, a format familiar to the users of microarrays for gene expression, the earliest high-throughput molecular chips.

The computation of statistical significance is an extremely common task, which has challenged researchers in particular for the computation of the corrections necessary in case of multiple hypotheses tested at one time. Given the ubiquitous nature of the operation and of the data that require it, it is necessary to provide formulas for the achievement of this goal which encompass the necessity to have large computing facilities only seldom available in biological lab, which may nevertheless require these computations.

Our package deals with this issue, and o□ers p-value based statistical significance scores, focusing on minimizing the number of permutations. Our approach allows to perform with enhanced and e□cient statistical approaches two very common operations (computation of similarity among genes and computation of similarity of genes with other traits) using correlation and mutual information.

BUS Vignette

Yin Jin, Hesen Peng, Lei Wang, Raffaele Fronza, Yuanhua Liu and Christine Nardini

1 Introduction

GOAL: The BUS package allows the computation of two types of similarities (correlation [Sokal, 2003] and mutual information [Cover, 2001]) for two different goals: (i) identification of the similarity among the activity of molecules sampled across different experiments (we name this option Unsupervised, U), (ii) identification of the similarity between such molecules and other types of information (clinical, anagraphical, etc, we name this option supervised, S).

Unsupervised Option. The computation applies to data in tabular form (MxN) where rows represents different molecules (M), columns represents experiments or samples (N) and the content of the tables' cells the abundance of the molecule in the sample. Microarray experiments are the data of choice for this application, but the method can be applied to any data in the appropriate format (miRNA arrays, RNA-seq data, etc.). The results are in the form of an MxM adjacency matrix, where each cell represents the association computed among the corresponding molecules. This matrix has associated also a p-value matrix and a corrected p-value matrix (see below for details). Based on the cutoff selected, the adjacency matrix can be trimmed and lead to a predicted network of statistically significant interactions (`pred.network`). This output can be used as-is to represent a gene association network ([Margolin, 2004, Basso, 2005]), or can be further elaborated to cluster genes based on a shared degree of similarity (hence the Unsupervised label). Mutual information (from now on MI) is computed using the minet package [Meyer, 2008], all the options can be found in the corresponding vignette. Here argument `net.trim` decides which function (mrnet/clr/aracne) in MINET package is used to give the similarity based on mutual information matrix. Correlation is computed using the R built-in `cor` function.

Supervised Option. For the S option a second dataset is necessary, a TxN table, where T represents the number of external traits of interest. The result is an association MxT table where each cell indicates the association between the molecule and the external trait. Mutual information is computed according to the empirical method proposed in MINET package. It is implemented with a external c function. This matrix has associated also a p-value matrix and a corrected p-value matrix (see below for details). As this can be used to associate samples to clinical classes we call this option Supervised (this type of approach was used in [Diehn, 2008]).

Statistical Significance. The package offers the possibility to evaluate the statistical significance of the computed similarity measures in two steps, a summary of the options is given in Table 1.

Option	p-value		
	single		multiple
	ρ	MI	MI
S	Exact	*beta* distribution	permutations (3 options)
U			permutations

Table 1. Summary of the available options for statistical validation in BUS. ρ indicates correlation.

First, it allows the computation of the "single" p-value, i.e. the p-value relevant for the assessment of the statistical significance of the similarity of a given gene as if it was the only one tested.

For correlation this relies on the R built-in cor.test and it then computes the exact p-value.

For MI it is obtained from permutations and this method estimates the extreme p-values (close to 0) by fitting a beta distribution, whose analytical expression is obtained by the estimate of 2 shape parameters ($\hat{\alpha}$ and $\hat{\beta}$) using the method of the moments.

Second, for the p-value of MI, correction for multiple hypothesis testing is computed based on permutations. 3 types of corrections are offered:

- S analysis option `method.permut` = 1 correction for multiple traits tested
- S analysis option `method.permut` = 2 correction for multiple genes tested
- S analysis option `method.permut` = 3 correction for both traits and genes

Missing Data Treatment. Data are pre-processed to cope with missing information (both in the MxN and in the TxN table) using (smooth) bootstrapping [Silverman, 1987].

The main function BUS has arguments for:

- the type of analysis (supervised/unsupervised)
- the distance metric (correlation/MI)
- the correction types for statistical significance on multiple hypothesis testing based on permutations (genes, traits or both)

Expected computation times. In the unsupervised case, the anticipated time for a 50*12 matrix (gene expression data) is 30 seconds when running on an ordinary personal computer (with 1G memory). While in the supervised case, with 50*12 gene expression data and trait data involved, it is 2 minute when correction of both genes and traits is considered.

The functions' dependencies scheme of the BUS package is illustrated below.

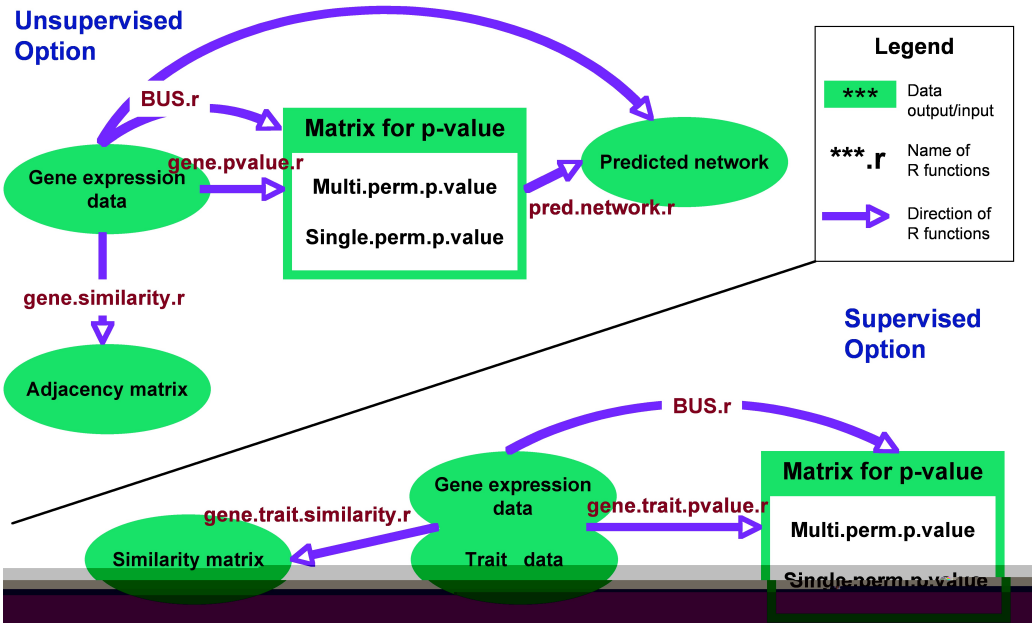


Figure 1. functions scheme

Functions Description

`BUS`: A wrapper function to compute (i) the similarity matrix (using correlation/MI as metric) and the single p-value matrix (each element is the p-value under the null hypothesis that the related row gene and column gene

have no interaction), corrected p-values matrix (different levels of dependency are considered) and the predicted network matrix (predicted gene network, this output is effective for option U)

`gene.similarity`: Function for the computation of the adjacency matrix in the Unsupervised option (using correlation/MI as metric)

`gene.trait.similarity`: Function for the computation of the similarity matrix in the Supervised option (using correlation/MI as metric)

`gene.pvalue`: Function for the computation of the p-value matrix for the Unsupervised option. Single p-value (each element is the p-value under the null hypothesis that the related row gene and column gene have no interaction) is computed thanks to: (i) for MI the distribution identified by the P permutation values identified for each gene, with extreme p-values computed fitting a beta distribution; for correlation using the exact distribution provided by the built-in R cor function (single.perm.p.value). Corrected p-value is computed thanks to the distribution identified by the p permutation values across all genes (multi.perm.p.value). When correlation is used as matric, only exact p-value is output.

`gene.trait.pvalue`: Function for the computation of the p-value matrix for the Supervised option. Single p-value (each element is the p-value under the null hypothesis that the related row gene and column trait have no interaction) is computed thanks to: (i) for MI the distribution identified by the P permutation values identified for each gene, with extreme p-values computed fitting a beta distribution; for correlation using the exact distribution provided by the built-in R `cor` function (single.perm.p.value). Corrected p-value is computed thanks to the distribution identified by the P permutation values across all genes (multi.perm.p.value); (ii) the distribution identified by the P permutation values across all traits; (iii) the distribution identified by the P permutation values across all genes and traits.

`pred.network`: Function to predict the network from the selected corrected p-value matrix, only for the Unsupervised option.

2 BUS Usage

```
> library(BUS)
> library(minet)
> data(copasi)
> mat = as.matrix(copasi)[1:5, ]
> rownames(mat) <- paste("G", 1:nrow(mat), sep = "")
> BUS(EXP = mat, measure = "MI", n.replica = 400,
+     net.trim = "aracne", thresh = 0.05, nflag = 1)

$similarity
          G1        G2        G3 G4        G5
G1 1.0000000 0.4682614 0.5126417  1 0.0000000
G2 0.4682614 1.0000000 0.0000000  0 0.7451189
G3 0.5126417 0.0000000 1.0000000  0 0.6168879
G4 1.0000000 0.0000000 0.0000000  1 0.0000000
G5 0.0000000 0.7451189 0.6168879  0 1.0000000

$single.perm.p.value
       G1     G2     G3     G4     G5
G1 0.0000 0.2075 0.2450 0.0000 0.5050
G2 0.2075 0.0000 0.4650 0.4775 0.1675
G3 0.2450 0.4650 0.0000 0.4500 0.1575
G4 0.0000 0.4775 0.4500 0.0000 0.4825
G5 0.5050 0.1675 0.1575 0.4825 0.0000

$multi.perm.p.value
          G1         G2        G3        G4         G5
G1 0.0000000 0.14494265 0.1305556 0.0000000 0.40182009
```

```
G2 0.1449426 0.00000000 0.3997930 0.4031812 0.04763561
G3 0.1305556 0.39979303 0.0000000 0.4017271 0.09482459
G4 0.0000000 0.40318119 0.4017271 0.0000000 0.41155989
G5 0.4018201 0.04763561 0.0948246 0.4115599 0.00000000

$net.pred.permut
   G1        G2 G3 G4        G5
G1  1 0.0000000  0  1 0.0000000
G2  0 1.0000000  0  0 0.7451189
G3  0 0.0000000  1  0 0.0000000
G4  1 0.0000000  0  1 0.0000000
G5  0 0.7451189  0  0 1.0000000
```

The arguments to the `BUS` function here are

- `EXP`, a matrix for gene expression data.
- `measure`, metric used to calculate similarity. There are two choices, MI and corr. We use MI here, applying the MINET package to output the similarity matrix with option of `aracne`.
- `method.permut`, a flag to indicate which method is used to correct permutation p-values. Here a default value (2) is used.
- `n.replica`, number of permutations: default value is 400, for optimal precision in p-value computation.
- `net.trim`, method chosen to trim the network. Here aracne method is applied, where the least significant edge in each triplet is removed.
- `threshold`, threshold, according to which significant association between genes are selected to construct the predicted network. This option is acutually used in function `pred.network` for predicted network from p-value matrix.
- `nflag`, a flag for the type of analysis. If Supervised `nflag`=2, if Unsupervised `nflag`=1. Here an Unsupervised option is considered.

The copasi dataset is taken from Copasi2 (Complex Pathway Simulator), a software for simulation and analysis of biochemical networks. The system generates random artificial gene networks according to well-defined topological and kinetic properties. These are used to run in silico experiments simulating real laboratory micro-array experiments. Noise with controlled properties is added to the simulation results several times emulating measurement replicates, before expression ratios are calculated. This series consists of 150 artificial gene networks. Each network consists of 100 genes with a total of 200 gene interactions (on average each gene has 2 modulators). All networks are composed of genes with similar kinetics, the only difference between networks is how the gene interactions are organized (i.e. which genes induce and repress which other genes). The networks belong to three major groups according to their topologies: RND stands for randomized network, SF for scale-free(many edges among few nodes) and SW for small world (edges exist between adjacent nodes). The data given in the package is an RND data. Actually, only first of five rows in the gene expression data is used to calculate to save the space here.

Explain the results:

- `similarity`: the matrix for mutual information.
- `single.perm.p.value`: the single p-value matrix, i.e. the p-value matrix obtained by the simple purmutation method. We can see it is a 5*5 matrix here as we only use data for 5 genes.
- `multi.perm.p.value`: the corrected permutation p-value matrix, i.e. the p-value matrix obtained via corrected permutation method.

- `net.pred.permut`: the network predicted based on the corrected permutation p-value matrix. This network is based on multi-hypothesis-corrected p-values.

This is an Unsupervised case. We could see that a lower values in `single.perm.p.value/multi.perm.p.value` or a higher values in `net.pred.permut` indicate a strong link between the row and column genes. The value 0 in the p-value matrix or 1 in network matrix respectively infers a strong link.

```
> data(tumors.mRNA)
> exp <- as.matrix(tumors.mRNA)[11:15, ]
> rownames(exp) <- rownames(tumors.mRNA)[11:15]
> data(tumors.miRNA)
> trait <- as.matrix(tumors.miRNA)[11:15, ]
> rownames(trait) <- rownames(tumors.miRNA)[11:15]
> BUS(EXP = exp, trait = trait, measure = "MI",
+     nflag = 2)

$similarity
            hsa-mir-132 hsa-mir-133a hsa-mir-135a
200017_at     0.6000000    0.4754888    0.4754888
200018_at     0.2000000    0.5509775    0.4754888
200022_at     0.4754888    0.5509775    0.0000000
200023_s_at   0.0000000    1.0000000    0.0000000
200024_at     0.4754888    0.5509775    0.0000000
            hsa-mir-135b hsa-mir-139
200017_at      0.4754888   1.0000000
200018_at      0.4754888   0.4754888
200022_at      0.0000000   0.4754888
200023_s_at    0.0000000   0.4754888
200024_at      0.4754888   0.4754888

$single.perm.p.value
            hsa-mir-132 hsa-mir-133a hsa-mir-135a
200017_at        0.2125       0.3700       0.3725
200018_at        0.6175       0.2975       0.4025
200022_at        0.3850       0.3250       0.6700
200023_s_at      0.7225       0.0000       0.7350
200024_at        0.3675       0.3200       0.6900
            hsa-mir-135b hsa-mir-139
200017_at         0.3400      0.0000
200018_at         0.4075      0.3600
200022_at         0.7350      0.3450
200023_s_at       0.7075      0.3850
200024_at         0.3875      0.3575

$multi.perm.p.value
            hsa-mir-132 hsa-mir-133a hsa-mir-135a
200017_at        0.2435       0.3605       0.3730
200018_at        0.6210       0.3175       0.3730
200022_at        0.3780       0.3175       0.7005
200023_s_at      0.6955       0.0000       0.7005
200024_at        0.3780       0.3175       0.7005
            hsa-mir-135b hsa-mir-139
200017_at         0.3805       0.000
200018_at         0.3805       0.357
```

```
200022_at       0.7060      0.357
200023_s_at     0.7060      0.357
200024_at       0.3805      0.357
```

Here is a Supervised case, we use the tumor dataset from [Liu, 2007], the mRNA data as gene expression data and miRNA data as trait data. Gene expression data were obtained by microarray from human brain tumors, while miRNA data were obtained by RT-PCR. 12 brain tumors at different levels are analyzed for both mRNA and miRNA levels to study the correlation of any mRNA-miRNA pairs. Outputs are similar like that in the unsupervised case except the predicted network.

References

[Sokal, 2003] R.R.Sokal and F.J.Rohlf. *Biometry.* Freeman, New York, 2003.

[Cover, 2001] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley and Sons, 2001.

[Margolin, 2004] A. A. Margolin, I. Nemenman, K. Basso, U. Klein, C. Wiggins, G. Stolovitzky, R. Dalla. Favera, and A. Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, 2004.

[Basso, 2005] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human b cells. *Nat Genet*, 37(4):382–390, Apr 2005.

[Meyer, 2008] P. E. Meyer, F Lafitte, and G. Bontempi. minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9:461–461, 2008.

[Diehn, 2008] M. Diehn, C. Nardini, D. S. Wang, S. McGovern, M. Jayaraman, Y. Liang, K. Aldape, S. Cha, and M. D. Kuo. Identification of non-invasive imaging surrogates for brain tumor gene expression modules. *Proc. Natl. Acad. Sci.*, 105(13):5213–5218, 2008.

[Silverman, 1987] B. W. Silverman and G. A. Young. The bootstrap: To smooth or not to smooth? *Biometrika*, 74(3):469–479, 1987.

[Liu, 2007] T. Liu, T. Papagiannakopoulos, K. Puskar, S. Qi, F. Santiago, W. Clay, K. Lao, Y. Lee, S. F. Nelson, H. I. Kornblum, F. Doyle, L. Petzold, B. Shraiman, and K. S. Kosik. Detection of a microRNA signal in an in vivo expression set of mRNAs. *Plos One*, 2(8): e804, 2007.

Bibliography

[ABC+10] Alan L Archibald, Lars Bolund, Carol Churcher, Merete Fredholm, Martien A M Groenen, Barbara Harlizius, Kyung-Tai Lee, Denis Milan, Jane Rogers, Max F Rothschild, Hirohide Uenishi, Jun Wang, Lawrence B Schook, and . Pig genome sequence–analysis and publication strategy. *BMC Genomics*, 11:438, 2010.

[ACA+07] Alessandro Aiuti, Barbara Cassani, Grazia Andolfi, Massimiliano Mirolo, Luca Biasco, Alessandra Recchia, Fabrizia Urbinati, Cristina Valacca, Samantha Scaramuzza, Memet Aker, Shimon Slavin, Matteo Cazzola, Daniela Sartori, Alessandro Ambrosi, Clelia Di Serio, Maria Grazia Roncarolo, Fulvio Mavilio, and Claudio Bordignon. Multilineage hematopoietic reconstitution without clonal selection in ada-scid patients treated with stem cell gene therapy. *J Clin Invest*, 117(8):2233–40, Aug 2007.

[AFN96] D B Allison, M S Faith, and J S Nathan. Risch's lambda values for human obesity. *Int J Obes Relat Metab Disord*, 20(11):990–9, Nov 1996.

[AKK+96] D B Allison, J Kaprio, M Korkeila, M Koskenvuo, M C Neale, and K Hayakawa. The heritability of body mass index among an international sample of monozygotic twins reared apart. *Int J Obes Relat Metab Disord*, 20(6):501–6, Jun 1996.

[Amb04] Victor Ambros. The functions of animal microrna. *Nature*, 431(7006):350–5, Sep 2004.

[AMS+97] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Sep 1997.

[ATPP06a] A C Ahn, M Tewari, C-S Poon, and R S Phillips. The clinical applications of a systems approach. *PLoS Medicine*, 3(7):e209, 2006.

[ATPP06b] A C Ahn, M Tewari, C-S Poon, and R S Phillips. The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Medicine*, 3(6):e208, 2006.

[AZFD05] W R Atchley, J Zhao, A D Fernandes, and T Drüke. Solving the protein sequence metric problem. *Proc Natl Acad Sci U S A*, 102(18):6395–6400, May 2005.

[Bar04] David P Bartel. Microrna: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, Jan 2004.

[BAW+05] Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole

Redaschi, and Lai-Su L Yeh. The universal protein resource (uniprot). *Nucleic Acids Res*, 33(Database issue):D154–9, Jan 2005.

[BGA+07] O A Burmistrova, A Y Goltsov, L I Abramova, V G Kaleda, V A Orlova, and E I Rogaev. Microrna in schizophrenia: genetic and expression analysis of mir-130b (22q11). *Biochemistry (Mosc)*, 72(5):578–82, May 2007.

[BGK08] Nicolas Bousette, Anthony O Gramolini, and Thomas Kislinger. Proteomics-based investigations of animal models of disease. *Proteomics Clin Appl*, 2(5):638–53, May 2008.

[BKM+06] Christopher Baum, Olga Kustikova, Ute Modlich, Zhixiong Li, and Boris Fehse. Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors. *Hum Gene Ther*, 17(3):253–63, Mar 2006.

[Bro99] A J Brookes. The essence of snps. *Gene*, 234(2):177–86, Jul 1999.

[BSS+10] Kaan Boztug, Manfred Schmidt, Adrian Schwarzer, Pinaki P Banerjee, Inés Avedillo Díez, Ricardo A Dewey, Marie Böhm, Ali Nowrouzi, Claudia R Ball, Hanno Glimm, Sonja Naundorf, Klaus Kühlcke, Rainer Blasczyk, Irina Kondratenko, László Maródi, Jordan S Orange, Christof von Kalle, and Christoph Klein. Stem-cell gene therapy for the wiskott-aldrich syndrome. *N Engl J Med*, 363(20):1918–27, Nov 2010.

[But02] A. Butte. The use and analysis of microarray data. *Nature Reviews Drug Discovery*, 1:951–960, 2002.

[BZ09] Ian S Blagbrough and Chiara Zara. Animal models for target diseases in gene therapy–using dna and sirna delivery strategies. *Pharm Res*, 26(1):1–18, Jan 2009.

[CCLHBAF05] Marina Cavazzana-Calvo, Chantal Lagresle, Salima Hacein-Bey-Abina, and Alain Fischer. Gene therapy for severe combined immunodeficiency. *Annu Rev Med*, 56:585–602, 2005.

[CCT99] A.M. Costantini, C. Cannella, and G. Tomassi. *Fondamenti di nutrizione umana*. Il Pensiero Scientifico, 1999.

[CGP+06a] A P Crijns, F Gerbens, A E Plantinga, G J Meersma, S de Jong, R M Hofstra, E G de Vries, A G van der Zee, G H de Bock, and G J te Meerman. *BMC Genomics*, 7:232–232, 2006.

[CGP+06b] Anne P G Crijns, Frans Gerbens, A Edo D Plantinga, Gert Jan Meersma, Steven de Jong, Robert M W Hofstra, Elisabeth G E de Vries, Ate G J van der Zee, Geertruida H de Bock, and Gerard J te Meerman. A biological question and a balanced (orthogonal) design: the ingredients to e□ciently analyze two-color microarrays with confirmatory factor analysis. *BMC Genomics*, 7:232, 2006.

[CHBAB+09] Nathalie Cartier, Salima Hacein-Bey-Abina, Cynthia C Bartholomae, Gabor Veres, Manfred Schmidt, Ina Kutschera, Michel Vidaud, Ulrich Abel, Liliane Dal-Cortivo, Laure Caccavelli, Nizar Mahlaoui, Véronique Kiermer, Denice Mittelstaedt, Céline Bellesme, Najiba Lahlou, François Lefrère, Stéphane Blanche, Muriel Audit, Emmanuel Payen, Philippe Leboulch, Bruno l'Homme, Pierre Bougnères, Christof Von Kalle, Alain Fischer, Marina Cavazzana-Calvo, and Patrick Aubourg. Hematopoietic stem cell gene therapy with a lentiviral vector in x-linked adrenoleukodystrophy. *Science*, 326(5954):818–23, Nov 2009.

[Con01] The Gene Ontology Consortium. Creating the gene ontology resource: Design and implementation. *Genome Res.*, 11(8):1425–1433, 2001.

[DHBAS+07] Annette Deichmann, Salima Hacein-Bey-Abina, Manfred Schmidt, Alexandrine Garrigue, Martijn H Brugman, Jingqiong Hu, Hanno Glimm, Gabor Gyapay, Bernard Prum, Christopher C Fraser, Nicolas Fischer, Kerstin Schwarzwaelder, Maria-Luise Siegler, Dick de Ridder, Karin Pike-Overzet, Steven J Howe, Adrian J Thrasher, Gerard Wagemaker, Ulrich Abel, Frank J T Staal, Eric Delabesse, Jean-Luc Villeval, Bruce Aronow, Christophe Hue, Claudia Prinz, Manuela Wissler, Chuck Klanke, Jean Weissenbach, Ian Alexander, Alain Fischer, Christof von Kalle, and Marina Cavazzana-Calvo. Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in scid-x1 gene therapy. *J Clin Invest*, 117(8):2225–32, Aug 2007.

[DSH+03] Glynn Jr Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Cli□ord Lane, and Richard A Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003.

[ESBB98] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95(25):14863–14868, 1998.

[FAW+95] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, J M Merrick, and et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, Jul 1995.

[FCC+09] Valerio Fulci, Teresa Colombo, Sabina Chiaretti, Monica Messina, Franca Citarella, Simona Tavolaro, Anna Guarini, Robin Foà, and Giuseppe Macino. Characterization of b- and t-lineage acute lymphoblastic leukemia by integrated analysis of microrna and mrna expression profiles. *Genes Chromosomes Cancer*, 48(12):1069–82, Dec 2009.

[FCD+00] T S Furey, N Cristianini, N Du□y, D W Bednarski, M Schummer, and D Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–14, Oct 2000.

[FFS+02] S C Fahrenkrug, B A Freking, T P L Smith, G A Rohrer, and J W Keele. Single nucleotide polymorphism (snp) discovery in porcine expressed genes. *Anim Genet*, 33(3):186–95, Jun 2002.

[FG55] I H FEIGIN and S W GROSS. Sarcoma arising in glioblastoma of the brain. *Am J Pathol*, 31(4):633–53, Jul-Aug 1955.

[FGW+95] C M Fraser, J D Gocayne, O White, M D Adams, R A Clayton, R D Fleischmann, C J Bult, A R Kerlavage, G Sutton, J M Kelley, R D Fritchman, J F Weidman, K V Small, M Sandusky, J Fuhrmann, D Nguyen, T R Utterback, D M Saudek, C A Phillips, J M Merrick, J F Tomb, B A Dougherty, K F Bott, P C Hu, T S Lucier, S N Peterson, H O Smith, C A 3rd Hutchison, and J C Venter. The minimal gene complement of mycoplasma genitalium. *Science*, 270(5235):397–403, Oct 1995.

[FO06] Sadaf Farooqi and Stephen O'Rahilly. Genetics of obesity in humans. *Endocr Rev*, 27(7):710–18, Dec 2006.

[Fri98] Nir Friedman. The bayesian structural em algorithm. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 98:129–138, 1998.

[GAG98] D Gordon, C Abajian, and P Green. Consed: a graphical tool for sequence finishing. *Genome Res*, 8(3):195–202, Mar 1998.

[Gel96] H Geldermann. [analysis of gene e□ects on performance characteristics]. *Dtsch Tierarztl Wochenschr*, 103(10):378–83, Oct 1996.

[GEP+09] Richard Gabriel, Ralph Eckenberg, Anna Paruzynski, Cynthia C Bartholomae, Ali Nowrouzi, Anne Arens, Steven J Howe, Alessandra Recchia, Claudia Cattoglio, Wei Wang, Katrin Faber, Kerstin Schwarzwaelder, Romy Kirsten, Annette Deichmann, Claudia R Ball, Kamaljit S Balaggan, Rafael J Yáñez-Muñoz, Robin R Ali, H Bobby Gaspar, Luca Biasco, Alessandro Aiuti, Daniela Cesana, Eugenio Montini, Luigi Naldini, Odile Cohen-Haguenauer, Fulvio Mavilio, Adrian J Thrasher, Hanno Glimm, Christof von Kalle, William Saurin, and Manfred Schmidt. Comprehensive genomic access to vector integration in clinical gene therapy. *Nat Med*, 15(12):1431–6, Dec 2009.

[GM70] A J Gibbs and G A McIntyre. The diagram, a method for comparing sequences. its use with amino acid and nucleotide sequences. *Eur J Biochem*, 16(1):1–11, Sep 1970.

[GN08] C. Guiducci and C. Nardini. High parallelism, portability and broad accessibility: Technologies for genomics. *ACM J. Emerg. Technol. Comput. Syst.*, 4(1):Article 3, 2008.

[GNT10] Jeremy Goecks, Anton Nekrutenko, James Taylor, and . Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[HBAVKS+03a] S Hacein-Bey-Abina, C Von Kalle, M Schmidt, M P McCormack, N Wulffraat, P Leboulch, A Lim, C S Osborne, R Pawliuk, E Morillon, R Sorensen, A Forster, P Fraser, J I Cohen, G de Saint Basile, I Alexander, U Wintergerst, T Frebourg, A Aurias, D Stoppa-Lyonnet, S Romana, I Radford-Weiss, F Gross, F Valensi, E Delabesse, E Macintyre, F Sigaux, J Soulier, L E Leiva, M Wissler, C Prinz, T H Rabbitts, F Le Deist, A Fischer, and M Cavazzana-Calvo. Lmo2-associated clonal t cell proliferation in two patients after gene therapy for scid-x1. *Science*, 302(5644):415–9, Oct 2003.

[HBAvKS+03b] Salima Hacein-Bey-Abina, Christof von Kalle, Manfred Schmidt, Françoise Le Deist, Nicolas Wul□raat, Elisabeth McIntyre, Isabelle Radford, Jean-Luc Villeval, Christopher C Fraser, Marina Cavazzana-Calvo, and Alain Fischer. A serious adverse event after successful gene therapy for x-linked severe combined immunodeficiency. *N Engl J Med*, 348(3):255–6, Jan 2003.

[Hei87] G. Heijne. *Sequence analysis in molecular biology: treasure trove or trivial pursuit*. Academic Press, 1987.

[HG93] R Holliday and G W Grigg. Dna methylation and mutation. *Mutat Res*, 285(1):61–7, Jan 1993.

[HHP79] K A Houpt, T R Houpt, and W G Pond. The pig as a model for the study of obesity and of control of food intake: a review. *Yale J Biol Med*, 52(3):307–29, May-Jun 1979.

[HJ08] R A Holt and S J Jones. The new paradigm of flow cell sequencing. *Genome Res*, 18(6):839–846, Jun 2008.

[Hoc05] J. F. Hocquette. Where are we in genomics? *Journal of Physiology and Pharmacology*, 56(3):37–70, 2005.

[HSL09] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[JW02] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle River, NJ, (2002).

[KHS09] V Narry Kim, Jinju Han, and Mikiko C Siomi. Biogenesis of small rnas in animals. *Nat Rev Mol Cell Biol*, 10(2):126–39, Feb 2009.

[KKK+09] Hindrik Hd Kerstens, Sonja Kollers, Arun Kommadath, Marisol Del Rosario, Bert Dibbits, Sylvia M Kinders, Richard P Crooijmans, and Martien Am Groenen. Mining for single nucleotide polymorphisms in pig genome sequence data. *BMC Genomics*, 10:4, 2009.

[KLVS+10] A Kasim, D Lin, S Van Sanden, D-A Clevert, L Bijnens, H Göhlmann, D Amaratunga, S Hochreiter, Z Shkedy, and W Talloen. Informative or noninformative calls for gene expression: A latent variable approach. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 4, 2010.

[KMPA01] J M Kijas, M Moller, G Plastow, and L Andersson. A frameshift mutation in mc1r and a high frequency of somatic reversions cause black spotting in pigs. *Genetics*, 158(2):779–85, Jun 2001.

[KN01] L Kruglyak and D A Nickerson. Variation is the spice of life. *Nat Genet*, 27(3):234–6, Mar 2001.

[Koh08] D B Kohn. Gene therapy for childhood immunological diseases. *Bone Marrow Transplant*, 41(2):199–205, Jan 2008.

[KTF+86] M Kishikawa, N Tsuda, H Fujii, I Nishimori, H Yokoyama, and M Kihara. Glioblastoma with sarcomatous component associated with myxoid change. a histochemical, immunohistochemical and electron microscopic study. *Acta Neuropathol*, 70(1):44–52, 1986.

[KWT+98] J M Kijas, R Wales, A Törnsten, P Chardon, M Moller, and L Andersson. Melanocortin receptor 1 (mc1r) mutations and coat color in pigs. *Genetics*, 150(3):1177–85, Nov 1998.

[LBB05] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, Jan 2005.

[LFG+07] Giovanni Lanza, Manuela Ferracin, Roberta Gafà, Angelo Veronese, Riccardo Spizzo, Flavia Pichiorri, Chang gong Liu, George A Calin, Carlo M Croce, and Massimo Negrini. mrna/microrna gene expression profile in microsatellite unstable colorectal cancer. *Mol Cancer*, 6:54, 2007.

[LHW+09] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and . The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, Aug 2009.

[LIT92] P Langley, W Iba, and K Thompson. An analysis of bayesian classifiers. *Proceedings of AAAI*, 92:223–228, 1992.

[LKM+83] P M Laskarzewski, P Khoury, J A Morrison, K Kelly, M J Mellies, and C J Glueck. Familial obesity and leanness. *Int J Obes*, 7(6):505–27, 1983.

[LLLR07] Sï¿œverine Landais, Sï¿œbastien Landry, Philippe Legault, and Eric Rassart. Oncogenic potential of the mir-106-363 cluster and its implication in human t-cell leukemia. *Cancer Res*, 67(12):5699–707, Jun 2007.

[LOF+08] Alexandros Laios, Sharon O'Toole, Richard Flavin, Cara Martin, Lynn Kelly, Martina Ring, Stephen P Finn, Ciara Barrett, Massimo Loda, Noreen Gleeson, Tom D'Arcy, Eamonn McGuinness, Orla Sheils, Brian Sheppard, and John O' Leary. Potential role of mir-9 and mir-223 in recurrent ovarian cancer. *Mol Cancer*, 7:35, 2008.

[LPP+07] Tsunglin Liu, Thales Papagiannakopoulos, Kathy Puskar, Shuping Qi, Fernando Santiago, William Clay, Kaiqin Lao, Yohan Lee, Stanley F Nelson, Harley I Kornblum, Frank Doyle, Linda Petzold, Boris Shraiman, and Kenneth S Kosik. Detection of a microrna signal in an in vivo expression set of mrnas. *PLoS One*, 2(8):e804, 2007.

[LRD08] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–8, Nov 2008.

[LS91] W H Li and L A Sadler. Low nucleotide diversity in man. *Genetics*, 129(2):513–23, Oct 1991.

[LSB+05a] J J Lozano, M Soler, R Bermudo, D Abia, P L Fernandez, T M Thomson, and A R Ortiz. Dual activation of pathways regulated by steroid receptors and peptide growth factors in primary prostate cancer revealed by factor analysis of microarray data. *BMC Genomics*, 6:109–109, 2005.

[LSB+05b] Juan Jose Lozano, Marta Soler, Raquel Bermudo, David Abia, Pedro L Fernandez, Timothy M Thomson, and Angel R Ortiz. Dual activation of pathways regulated by steroid receptors and peptide growth factors in primary prostate cancer revealed by factor analysis of microarray data. *BMC Genomics*, 6:109, 2005.

[LTTD+00] K Lindblad-Toh, D M Tanenbaum, M J Daly, E Winchester, W O Lui, A Villapakkam, S E Stanton, C Larsson, T J Hudson, B E Johnson, E S Lander, and M Meyerson. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat Biotechnol*, 18(9):1001–1005, Sep 2000.

[LTWD+00] K Lindblad-Toh, E Winchester, M J Daly, D G Wang, J N Hirschhorn, J P Laviolette, K Ardlie, D E Reich, E Robinson, P Sklar, N Shah, D Thomas, J B Fan, T Gingeras, J Warrington, N Patil, T J Hudson, and E S Lander. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet*, 24(4):381–6, Apr 2000.

[LXKC02] Cesar Llave, Zhixin Xie, Kristin D Kasschau, and James C Carrington. Cleavage of scarecrow-like mrna targets directed by a class of arabidopsis mirna. *Science*, 297(5589):2053–6, Sep 2002.

[Men08] Joshua T Mendell. miriad roles for the mir-17-92 cluster in development and disease. *Cell*, 133(2):217–22, Apr 2008.

[MG77] A M Maxam and W Gilbert. A new method for sequencing dna. *Proc Natl Acad Sci U S A*, 74(2):560–4, Feb 1977.

[MHH+05] T Mijalski, A Harder, T Halder, M Kersten, M Horsch, T M Strom, H V Liebscher, F Lottspeich, M H de Angelis, and J Beckers. Identification of

coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues. *Proc Natl Acad Sci U S A*, 102(24):8621–8626, Jun 2005.

[MLB08] P E Meyer, F Lafitte, and G Bontempi. minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9:461–461, 2008.

[MLE+03] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, 34(3):267–273, 2003.

[MM88] E W Myers and W Miller. Optimal alignments in linear space. *Comput Appl Biosci*, 4(1):11–7, Mar 1988.

[MNB+04] Adam A. Margolin, Ilya Nemenman, Katia Basso, Ulf Klein, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, 2004.

[MNE97] H H Maes, M C Neale, and L J Eaves. Genetic and environmental factors in relative body weight and human adiposity. *Behav Genet*, 27(4):325–51, Jul 1997.

[MO04] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[MTL02] Bin Ma, John Tromp, and Ming Li. Patternhunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–5, Mar 2002.

[MW89] P R Mueller and B Wold. In vivo footprinting of a muscle specific enhancer by ligation mediated pcr. *Science*, 246(4931):780–6, Nov 1989.

[NMB+03] Catherine L Nutt, D R Mani, Rebecca A Betensky, Pablo Tamayo, J Gregory Cairncross, Christine Ladd, Ute Pohl, Christian Hartmann, Margaret E McLaughlin, Tracy T Batchelor, Peter M Black, Andreas von Deimling, Scott L Pomeroy, Todd R Golub, and David N Louis. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, 63(7):1602–7, Apr 2003.

[NRT+07] N Neretti, D Remondini, M Tatar, J M Sedivy, M Pierini, D Mazzatti, J Powell, C Franceschi, and G C Castellani. Correlation analysis reveals the emergence of coherence in the gene expression dynamics following system perturbation. *BMC Bioinformatics*, 8 Suppl 1, 2007.

[NTT97] D A Nickerson, V O Tobe, and S L Taylor. Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*, 25(14):2745–51, Jul 1997.

[NW70] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53, Mar 1970.

[PAG+10] Anna Paruzynski, Anne Arens, Richard Gabriel, Cynthia C Bartholomae, Simone Scholz, Wei Wang, Stephan Wolf, Hanno Glimm, Manfred Schmidt, and Christof von Kalle. Genome-wide high-throughput integrome analyses by nrlam-pcr and next-generation sequencing. *Nat Protoc*, 5(8):1379–95, Aug 2010.

[PBK+10] S K Panguluri, S Bhatnagar, A Kumar, J J McCarthy, A K Srivastava, N G Cooper, R F Lundy, and A Kumar. Genomic profiling of messenger rnas and micrornas reveals potential mechanisms of tweak-induced skeletal muscle wasting in mice. *PLoS One*, 5(1), 2010.

[Pet02a] L E Peterson. Factor analysis of cluster-specific gene expression levels from cdna microarrays. *Comput Methods Programs Biomed*, 69(3):179–188, Nov 2002.

[Pet02b] Leif E Peterson. Factor analysis of cluster-specific gene expression levels from cdna microarrays. *Comput Methods Programs Biomed*, 69(3):179–88, Nov 2002.

[PKS+07] Oscar Persson, Morten Krogh, Lao H Saal, Elisabet Englund, Jian Liu, Ramon Parsons, Nils Mandahl, Ake Borg, Bengt Widegren, and Leif G Salford. Microarray analysis of gliomas reveals chromosomal position-associated gene expression patterns and identifies potential immunotherapy targets. *J Neurooncol*, 85(1):11–24, Oct 2007.

[PRS+09] Giorgos L Papadopoulos, Martin Reczko, Victor A Simossis, Praveen Sethupathy, and Artemis G Hatzigeorgiou. The database of experimentally supported targets: a functional update of tarbase. *Nucleic Acids Res*, 37(Database issue):D155–8, Jan 2009.

[PSH+07] Frank Panitz, Henrik Stengaard, Henrik Hornshøj, Jan Gorodkin, Jakob Hedegaard, Susanna Cirera, Bo Thomsen, Lone B Madsen, Anette Høj, Rikke K Vingborg, Bujie Zahn, Xuegang Wang, Xuefei Wang, Rasmus Wernersson, Claus B Jørgensen, Karsten Scheibye-Knudsen, Troels Arvin, Steen Lumholdt, Milena Sawera, Trine Green, Bente J Nielsen, Jakob H Havgaard, Søren Brunak, Merete Fredholm, and Christian Bendixen. Snp mining porcine ests with maviant, a novel tool for snp evaluation and annotation. *Bioinformatics*, 23(13):i387–91, Jul 2007.

[PW07] I Pournara and L Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8:61–61, 2007.

[Qua01] J. Quackenbush. Computational analysis of micorarray data. *Nat Rev Genet*, 2(6):418–427, 2001.

[RCA+09] Antonio M Ramos, Richard P M A Crooijmans, Nabeel A A□ara, Andreia J Amaral, Alan L Archibald, Jonathan E Beever, Christian Bendixen, Carol Churcher, Richard Clark, Patrick Dehais, Mark S Hansen, Jakob Hedegaard, Zhi-Liang Hu, Hindrik H Kerstens, Andy S Law, Hendrik-Jan Megens, Denis Milan, Danny J Nonneman, Gary A Rohrer, Max F Rothschild, Tim P L Smith, Robert D Schnabel, Curt P Van Tassell, Jeremy F Taylor, Ralph T Wiedmann, Lawrence B Schook, and Martien A M Groenen. Design of a high density snp genotyping assay in the pig using snps identified and characterized by next generation sequencing technology. *PLoS One*, 4(8):e6524, 2009.

[RGA03] David E Reich, Stacey B Gabriel, and David Altshuler. Quality and completeness of snp databases. *Nat Genet*, 33(4):457–8, Apr 2003.

[RKP+96] M Ronaghi, S Karamohamed, B Pettersson, M Uhlén, and P Nyrén. Real-time dna sequencing using detection of pyrophosphate release. *Anal Biochem*, 242(1):84–9, Nov 1996.

[RPK03] R Roehe, G S Plastow, and P W Knap. Quantitative and molecular genetic determination of protein and fat deposition. *Homo*, 54(2):119–31, 2003.

[RSS01] M Remm, C E Storm, and E L Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–52, Dec 2001.

[SCS+03] Manfred Schmidt, Denise A Carbonaro, Carsten Speckmann, Manuela Wissler, John Bohnsack, Melissa Elder, Bruce J Aronow, Jan A Nolta, Donald B Kohn, and Christof von Kalle. Clonality analysis after retroviral-mediated gene transfer to cd34+ cells from the cord blood of ada-deficient scid neonates. *Nat Med*, 9(4):463–8, Apr 2003.

[SFH86] A J Stunkard, T T Foch, and Z Hrubec. A twin study of human obesity. *JAMA*, 256(1):51–4, Jul 1986.

[SG08] Michael E Spurlock and Nicholas K Gabler. The development of porcine models of obesity and the metabolic syndrome. *J Nutr*, 138(2):397–402, Feb 2008.

[She01] G Sherlock. Analysis of large-scale gene expression data. *Brief Bioinform*, 2(4):350–62, Dec 2001.

[SHG] AFA Smit, R Hubley, and P Green. Repeatmasker open-3.2.9. Unpublished data.

[SHS+07] Kerstin Schwarzwaelder, Steven J Howe, Manfred Schmidt, Martijn H Brugman, Annette Deichmann, Hanno Glimm, Sonja Schmidt, Claudia Prinz, Manuela Wissler, Douglas J S King, Fang Zhang, Kathryn L Parsley, Kimberly C Gilmour, Joanna Sinclair, Jinhua Bayford, Rachel Peraj, Karin Pike-Overzet, Frank J T Staal, Dick de Ridder, Christine Kinnon, Ulrich Abel, Gerard Wagemaker, H Bobby Gaspar, Adrian J Thrasher, and Christof von Kalle. Gammaretrovirus-mediated correction of scid-x1 is associated with skewed vector integration site distribution in vivo. *J Clin Invest*, 117(8):2241–9, Aug 2007.

[SJ06] C Sabatti and G M James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–746, Mar 2006.

[SM10] Ariella Sasson and Todd P Michael. Filtering error from solid output. *Bioinformatics*, 26(6):849–50, Mar 2010.

[SNC77] F Sanger, S Nicklen, and A R Coulson. Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–7, Dec 1977.

[SR03] R. R. Sokal and F. J. Rohlf. *Biometry*. Freeman, New York, 2003.

[SSB+07] Manfred Schmidt, Kerstin Schwarzwaelder, Cynthia Bartholomae, Karim Zaoui, Claudia Ball, Ingo Pilz, Sandra Braun, Hanno Glimm, and Christof von Kalle. High-resolution insertion-site analysis by linear amplification-mediated pcr (lam-pcr). *Nat Methods*, 4(12):1051–7, Dec 2007.

[SSB+09] Manfred Schmidt, Kerstin Schwarzwaelder, Cynthia C Bartholomae, Hanno Glimm, and Christof von Kalle. Detection of retroviral integration sites by linear amplification-mediated pcr and tracking of individual integration clones in di□erent samples. *Methods Mol Biol*, 506:363–72, 2009.

[SSH+86] A J Stunkard, T I Sørensen, C Hanis, T W Teasdale, R Chakraborty, W J Schull, and F Schulsinger. An adoption study of human obesity. *N Engl J Med*, 314(4):193–8, Jan 1986.

[SWH+09] Daniel Summerer, Haiguo Wu, Bettina Haase, Yang Cheng, Nadine Schracke, Cord F Stähler, Mark S Chee, Peer F Stähler, and Markus Beier. Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. *Genome Res*, 19(9):1616–21, Sep 2009.

[SWS+01] R Sachidanandam, D Weissman, S C Schmidt, J M Kakol, L D Stein, G Marth, S Sherry, J C Mullikin, B J Mortimore, D L Willey, S E Hunt, C G Cole, P C Coggill, C M Rice, Z Ning, J Rogers, D R Bentley, P Y Kwok, E R Mardis, R T Yeh, B Schultz, L Cook, R Davenport, M Dante, L Fulton, L Hillier, R H Waterston, J D McPherson, B Gilman, S Scha□ner, W J Van Etten, D Reich, J Higgins, M J Daly, B Blumenstiel, J Baldwin, N Stange-Thomann, M C Zody, L Linton, E S Lander, D Altshuler, and . A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–33, Feb 2001.

[TCRR+02] Joachim Theilhaber, Timothy Connolly, Sergio Roman-Roman, Steven Bushnell, Amanda Jackson, Kathy Call, Teresa Garcia, and Roland Baron. Finding genes in the c2c12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome Res*, 12(1):165–76, Jan 2002.

[THU46] L L THURSTONE. A single plane method of rotation. *Psychometrika*, 11:71–9, Jun 1946.

[THU47] L L THURSTONE. Factorial analysis of body measurements. *Am J Phys Anthropol*, 5(1):15–28, Mar 1947.

[TLV+08] Jifeng Tang, Jack A M Leunissen, Roeland E Voorrips, C Gerard van der Linden, and Ben Vosman. Haplosnper: a web-based allele and snp detection tool. *BMC Genet*, 9:23, 2008.

[TRBZ03] Guiliang Tang, Brenda J Reinhart, David P Bartel, and Phillip D Zamore. A biochemical framework for rna silencing in plants. *Genes Dev*, 17(1):49–63, Jan 2003.

[TS96] ME Tumbleson and LB Schook. *Advences in Swine in Biomedical Research*. Plenum Press, New York, 1996.

[TTC01] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.*, 98(9):5116–5121, 2001.

[TVV+06] Jifeng Tang, Ben Vosman, Roeland E Voorrips, C Gerard van der Linden, and Jack A M Leunissen. Qualitysnp: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in est data from diploid and polyploid species. *BMC Bioinformatics*, 7:438, 2006.

[VCL+06] S Volinia, G A Calin, C G Liu, S Ambs, A Cimmino, F Petrocca, R Visone, M Iorio, C Roldo, M Ferracin, R L Prueitt, N Yanaihara, G Lanza, A Scarpa, A Vecchione, M Negrini, C C Harris, and C M Croce. A microrna expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A*, 103(7):2257–2261, Feb 2006.

[VR02] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[vZ85] D von Zerssen. Psychiatric syndromes from a clinical and a biostatistical point of view. *Psychopathology*, 18(2-3):88–97, 1985.

[WAC07] Hui Wang, Robert A Ach, and Bo Curry. Direct and sensitive mirna profiling from low-input total rna. *RNA*, 13(1):151–9, Jan 2007.

[WBM+10] Gary P Wang, Charles C Berry, Nirav Malani, Philippe Leboulch, Alain Fischer, Salima Hacein-Bey-Abina, Marina Cavazzana-Calvo, and Frederic D Bushman. Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human scid-x1 gene therapy trial. *Blood*, 115(22):4356–66, Jun 2010.

[WGS09] Z Wang, M Gerstein, and M Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[Wil08] A E Williams. Functional aspects of animal micrornas. *Cell Mol Life Sci*, 65(4):545–62, Feb 2008.

[WL09] Yu-Ping Wang and Kuo-Bin Li. Correlation of expression profiles between micrornas and mrna targets using nci-60 data. *BMC Genomics*, 10:218, 2009.

[WNK+07] Haoquan Wu, Joel R Neilson, Priti Kumar, Monika Manocha, Premlata Shankar, Phillip A Sharp, and N Manjunath. mirna profiling of naive, effector and memory cd8 t cells. *PLoS One*, 2(10):e1020, 2007.

[WSGL07] Yang Wang, Heidi M Stricker, Deming Gou, and Lin Liu. Microrna: past and present. *Front Biosci*, 12:2316–29, 2007.

[YCL+06] T P Yang, T Y Chang, C H Lin, M T Hsu, and H W Wang. Arrayfusion: a web application for multi-dimensional analysis of cgh, snp and microarray data. *Bioinformatics*, 22(21):2697–2698, Nov 2006.

[YiYB+08] Man Lung Yeung, Jun ichirou Yasunaga, Yamina Bennasser, Nelson Dusetti, David Harris, Nafees Ahmad, Masao Matsuoka, and Kuan-Teh Jeang. Roles for micrornas, mir-93 and mir-130b, and tumor protein 53-induced nuclear protein 1 tumor suppressor in cell growth dysregulation by human t-cell lymphotrophic virus 1. *Cancer Res*, 68(21):8976–85, Nov 2008.

[YKV+08] N Yang, S Kaur, S Volinia, J Greshock, H Lassus, K Hasegawa, S Liang, A Leminen, S Deng, L Smith, C N Johnstone, X M Chen, C G Liu, Q Huang, D Katsaros, G A Calin, B L Weber, R Bützow, C M Croce, G Coukos, and L Zhang. Microrna microarray identifies let-7i as a novel biomarker and therapeutic target in human epithelial ovarian cancer. *Cancer Res*, 68(24):10307–10314, Dec 2008.

[YWF+06] J Yao, S Weremowicz, B Feng, R C Gentleman, J R Marks, R Gelman, C Brennan, and K Polyak. Combined cdna array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res*, 66(8):4065–4078, Apr 2006.

[ZFHEB03] Zhongming Zhao, Yun-Xin Fu, David Hewett-Emmett, and Eric Boerwinkle. Investigating single nucleotide polymorphism (snp) density in the human genome and its implications for molecular evolution. *Gene*, 312:207–13, Jul 2003.

Index