UNIVERSITÀ DEGLI STUDI DI SASSARI

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE

Direttore della Scuola: Prof. Franca Deriu

INDIRIZZO IN GENETICA MEDICA, MALATTIE METABOLICHE E NUTRIGENOMICA

Responsabile di Indirizzo: Prof. Francesco Cucca

XXVI CICLO

ANALYSIS OF QUALITATIVE AND QUANTITATIVE EXPRESSED TRAITS IN THE SARDINIAN POPULATION USING NEXT GENERATION SEQUENCING

Direttore:

Prof. Franca Deriu

Tutor:

Prof. Francesco Cucca

Tesi di dottorato di:

Dott. Mauro Pala

Anno Accademico 2012 - 2013

Summary

1 Introduction

1.1 Aim and project overview

Understanding the genetic basis of natural variation in gene expression has become crucial to dissect molecular mechanisms of complex traits and diseases[1,2]. As a matter of fact, only a minority of genetic variants identified by genome-wide association studies (GWAS) maps to protein-coding regions and only a few of them overtly alter the protein structure. Thus, for the remaining variants it has been hypothesized that they may have a subtler role in gene expression regulation.

Studies conducted thus far to assess systematically gene expression were limited by several factors. First, they were mostly based on heterogeneous, often multi-ethnic, populations[3,4]. Association studies of any type, including eQTL, among admixed populations are challenging because of the presence of confounding factors due to population stratifications, and heterogeneity of effects due to differential Linkage Disequilibrium (LD) blocks. Second, typically there are not large-scale phenotypic data from the individuals assessed in these expression studies. Hence, the extrapolation of the consequence to specific phenotypes could only be indirect, through coincident associations of phenotypic and expression traits with the same variants. Third, in some cases the cells from which the RNA was isolated are from disease tissues, which could complicate inference in unaffected individuals. Finally, the majority of the studies conducted so far relied on PolyA(+) enrichment protocols.

Most of the human genome transcribes not only protein-coding genes but also a large number of non-coding RNAs (ncRNAs). Among them, long non-coding RNAs (RNA with more than 200 nucleotides) are attractive because they have been implicated in several biological processed and diseases. Because of they lack the PolyA tail, their genetic regulation is poorly characterized.

We started an expression Quantitative Trait Loci (eQTL) study with the aim to overcome all these limitations.

We mapped eQTL in 608 individuals, a comparable sample size respect to the recently published eQTL studies on European (922 individuals)[4] and European and Africans (462 individuals)[3]. All the 608 volounteers were enrolled within the cohort of SardiNIA project[5]. As we will describe in the next chapter, they are extensively phenotypically characterized. Furthermore, while previous studies were based on heterogeneous populations, we decided to focus only on one homogeneous population, the Sardinian founder population.

RNA samples were sequenced using NGS (RNA-seq) which represents the most advanced high-throughput technology for the gene expression characterization. Unfortunately, a unique experimental approach cannot be exhaustive to characterize all the RNA species. In fact, the set of transcripts that can be studied depends largely by the method used for RNA preparation and for the library construction. Four main RNA populations can be chosen for sequencing: total RNA, ribosomal depleted RNA, polyadenylated RNA (polyA(+)), and not polyadenilated RNA (polyA(-)). First of all, we were interested in establishing which, between PolyA(+) selection and rRNA depletion, would be the best strategy to characterize the transcriptome of the entire cohort. For this reason we performed a pilot eQTL study with a subset of individuals processed in parallel with the two methods.

This work is focused mainly on the pilot project, with the aim to establish which RNA enrichment approach would be the choice in the context of a first large eQTL study in the Sardinian population.

The first part of this work was focused on the comparison between the libraries composition in terms of (i) distribution of reads among exons, introns and inter-genic regions, and (ii) genes classes differentially covered.

In the second part, for both of the libraries preparations, we computed the expression levels of genes, isoforms and splicing sites, and then we mapped the

genetic variants associated with changes in the expression levels of protein-coding genes (eQTL), long non-coding RNAs (lincQTL), splice-sites ratios (sQTL) and isoforms ratios (isoQTL). Then, we compared the results between the lists of eQTL obtained from the PolyA(+) selection and from the rRNA depletion protocols.

In the last part some preliminary results about the main project are described.

The PolyA(+) selection resulted better performing so we decided to adopt the PolyA(+) selection for the whole cohort of 608 individuals. We will show some statistics about the eQTL in the whole dataset of 608, and some preliminary comparisons with eQTL derived from the pilot project and recently published studies.

All these analysis were performed using public available software, setting up several data analysis pipelines and implementing new software. In particular we will describe more deeply two of them, CountSeq, a tool for quality controls and expression level quantification, and COViewer, a tool for the visualization of multi-samples RNA-seq data.

1.2 The uniqueness of the Sardinian genome: a founder population

Sardinia is an island located in the middle of the Mediterranean Sea, about 200 km from Italy, has about 1.6 millions inhabitants and is considered a large founder population. A founder population is a modern group that expanded from an initial small cohort with modest in-migration. Sardinia was first settled ~7700 years ago, during the pre-Neolithic or Neolithic age[6,7]. Despite several invasions and some trade relations, the immigration level has been modest. The consequent genetic drift has differentiated the Sardinian genetic pool from the European and African populations. For this reason, the Sardinian genome presents several unique characteristics. Founder-effects have amplified certain variants while maintaining great uniformity in other DNA sequences, causing a high degree of homogeneity, interrelatedness and enrichment in genetic variants that in other populations are rare[8].

These skewed gene frequencies can be used to study rare, single-gene disorders but also complex diseases and traits. Like other founder populations, in Sardinia there are increased disease frequencies, specifically auto-immune diseases, like Type I diabetes (T1D) and Multiple Sclerosis (MS)[9].

Ogliastra, a region about 60,000 inhabitants, was chosen to start a study focused on risk factors and quality of life: the SardiNIA project. This project is a longitudinal study that phenotyped around 6000 individuals, males and females aged from 4-102 years[5], and 1,629 of them that have been also extensively genotyped and characterized for blood cell composition[10].

For all of these reasons, the SardinNIA project is an attractive setting for studying the genetic bases of complex trait and diseases. Under this context, we hypothesize that the characterization of the transcriptome will provide insights on the molecular mechanisms of these bases.

1.3 Dissecting the genetics of complex phenotypes with eQTL: A common strategy

In order to understand the genetic basis of the natural gene expression variation, numerous studies had associated gene expression levels with genetic polymorphism[2-4,11-16]. The general approach is to consider gene expression levels as quantitative traits and map them to the genome through association or linkage analysis. Those genetic variants that show association with the gene expression levels are called expression Quantitative Traits Loci (eQTL).

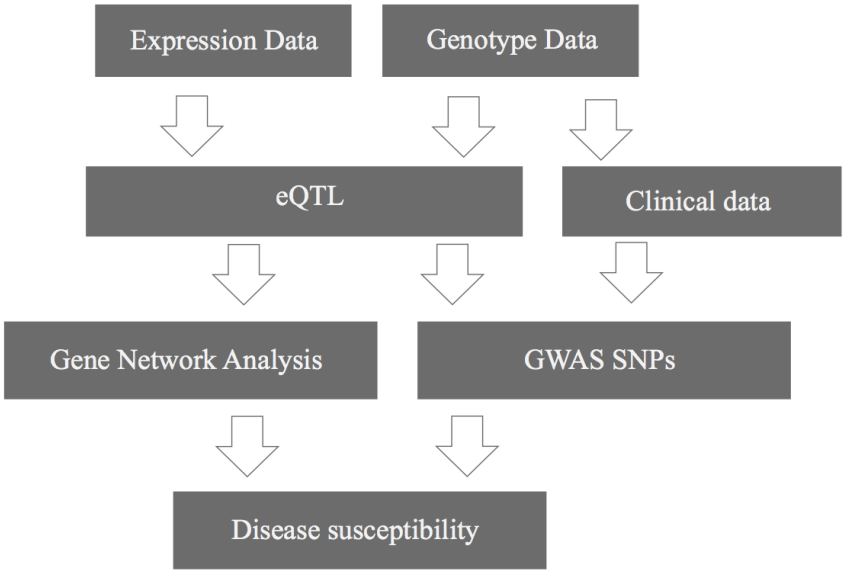

Figure 1. eQTL mapping and complex traits and diseases. eQTL mapping starts with the collection of expression levels measurements. Than association is performed with the genotype. Network analysis based on gene expression level correlation, can identify network modules and dissect the role of variants associated with complex traits and diseases.

The common workflow for eQTL mapping is illustrated in Figure 1. It starts with the measurement of the gene expression in a target cell or tissue from multiple individuals. The quantification of the expression levels are usually performed with high-throughput technologies, like microarrays or RNA-seq. eQTL are then mapped in the same manner as any other quantitative trait[17,18].

Several insights have been provided with the eQTL studies so far. Variation in gene expression levels and alternative splicing have shown to be highly heritable[19,20]. It has been hypothesized that this was mostly due to the fact that, respect to complex traits (like blood pressure or weight), there is a smaller number of molecular interactions between the regulatory variant and the expression level changes[2].

eQTL can be classified in *cis*-eQTL and *trans*-eQTL, based on the genomic distance respect to the target gene. Usually *cis*-eQTL are located within 1-2 Mb from the target gene, while *trans*-eQTL map to larger distances or different chromosomes. Due to their close proximity, *cis*-eQTL most likely control the expression level of nearby genes. *cis*-QTL are often found near the Transcriptional Start Site (TSS)[16,21], and has been estimated that 90% of them tends to fall within 15kb from the gene boudaries[16]. On the other end, *trans*-eQTL generally are variants that act in *trans*, and are more difficult to be identified with certainty because significant threshold must be set very high since all the regions of the genome must be tested.

Several publications reported polymorphisms associated not only with the expression levels, but also with the alternative splicing (sQTL)[16,22] and the isoforms usage[15]. sQTL have been largely studied and has been reported that have been subject to strong positive selection in recent human history and that are enriched for SNPs associated with autoimmune diseases, suggesting that these SNPs most likely act through alteration of splicing and polyadenylation site usage[22].

It has been found that SNPs located in splicing factors binding sites are enriched among sQTL, relatively to non-splice site intronic variants[16]. Furthermore, SNPs situated in the canonical splice sites are enriched in sQTL[16] and they are more associated with the first, second and last exon compared with any middle exon[15].

eQTL studies have also provided several insights about the role of loci

associated with complex traits and diseases[2,22-24]. *Cis*-regulated Vanin 1 (VNN1) gene was found to influence high-density lipoprotein cholesterol concentration[14]. APOE (apolipoprotein E) and MAPT (microtubule-associated protein tau) were found to play an important role in Alzheimer's disease[25]. ORMLD3 (ORM1-like 3) expression-level was found to be associated with a SNP associated with the risk of asthma[23].

eQTL approaches have been also integrated with gene expression and clinical data to infer causal relationships among gene expression traits and between expression and clinical traits[26]. These studies proposed an alternative approach to the common association method. Instead of identifying susceptibility genes directly affected by variations in DNA, they individuated gene networks that were perturbed by susceptibility loci and that in turn leaded to disease. In one of these studies[27], three genes, lipoprotein lipase (Lpl), lactamase b (Lactb) and protein phosphatase 1-like (Ppm1l), were validated as previously unknown obesity genes.

However, these results have been mostly possible because was used a tissue relevant to the interrogated complex trait. In fact, it is expected that functional variants may operate in a tissue-dependent manner[28] although the extent of this is still under debate.

Understanding such complex effects in the contest of gene expression is essential for the dissection of the molecular basis of complex traits and diseases and the achievement of this understanding could be dramatically improved or even uniquely obtained only with studies that combine genomic and transcriptomic data, with the availability of the relavant tissue and the clinical informations.

1.4 RNA-seq is the most advanced technology for studying RNA samples

In order to characterize the transcriptome we used the High-Throughput RNA Sequencing (RNA sequencing or RNA-seq). RNA-seq is a relatively recent developed technology and is the first sequencing-based method that is both quantitative and and high-throughput. For these reasons, respect to the existing techniques, in the context of whole-transcriptome studies, the RNA-seq offers a better balance between informativity and costs. Here we will briefly describe the technique and its advantages.

Data generation is based on the Next-Generation Sequencing technologies (NGS)[29] and the workflow is represented in Figure 1. Briefly, RNA samples are extracted from the tissues or cells, fractioned (such as PolyA(+) selection or ribosomal depletion), fragmented and converted to a library of cDNA with adaptors attached to one or both the ends. Each cDNA fragment, with or without amplification, is then attached to the flowcell and sequenced with NGS technologies to obtain short sequences from one end (single-end) or both ends (paired-end). One such sequence is called *read*. It is typically 50-100bp. In a typical RNA-seq experiment tens of millions of reads are obtained from a sample.

Data analysis can be summarized in three steps: assignment of reads to the reference genome, quality controls and expression quantification.
The assignment of reads to their genomic position can be achieved in two ways, with *de novo* assembling[30] or by the alignment to the reference genome[31]. The *de novo* assembling approach is mostly used to map genes in the absence of poorly annotated genomes. There is still not a unified consensus about the reliability of *de novo* assembling in the context of expression quantification thus, for most of

the RNA-seq applications, the mapping to the reference genome is preferred. A large number of aligners is now available for RNA-seq[31]. They differ from common aligners because they are able to map reads that span exon-exon junctions (spliced reads, SR) and to determine exon-intron boundaries (Figure 2). Reads can map to unique loci of the genome (Unique mapping Reads, UR) or to multiple loci (Multiple mapped Reads or MR). MR can belong to transcribed repeated regions (rRNAs, tRNAs, snoRNAs) or to regions with high similarity in sequence. Enven if sofware that deal with such reads are available[32], MR are usually discarded for downstream analysis.

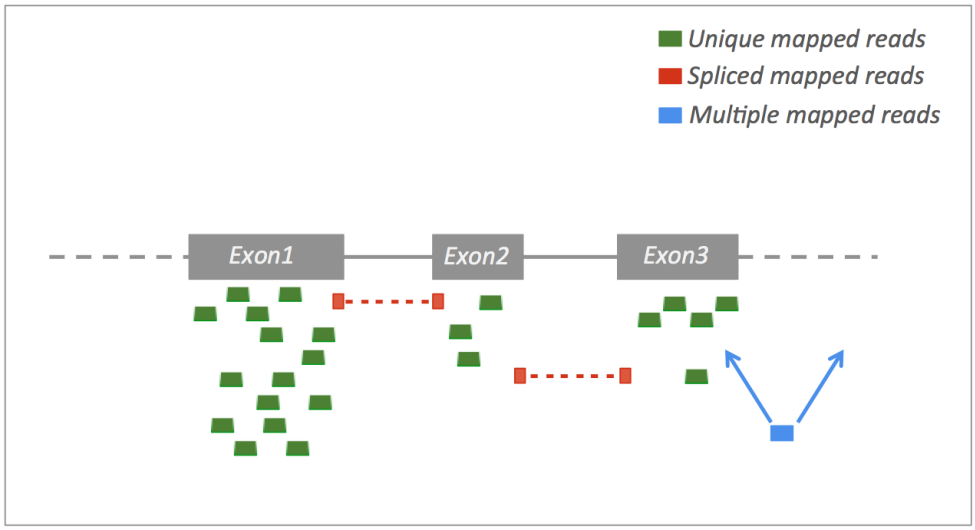

Figure 2. Classification of reads based on mapping features. Red dashed lines indicate reads spanning exon-exon junctions.

Quality controls are essential to recognize artifacts that can affect the subsequent quantification analysis. For example, the proportion of rRNAs in the samples is used to evaluate the RNA fractionation step (like PolyA(+) selection or ribosomal depletion). Other common quality metrics are computed at read-level (number of reads sequenced, reads mapped and reads with different levels of mapping qualities) or at base-level (GC-content, base quality distribution, base composition) to assess the sequencing quality and efficiency. Some examples are shown in Figure 3.

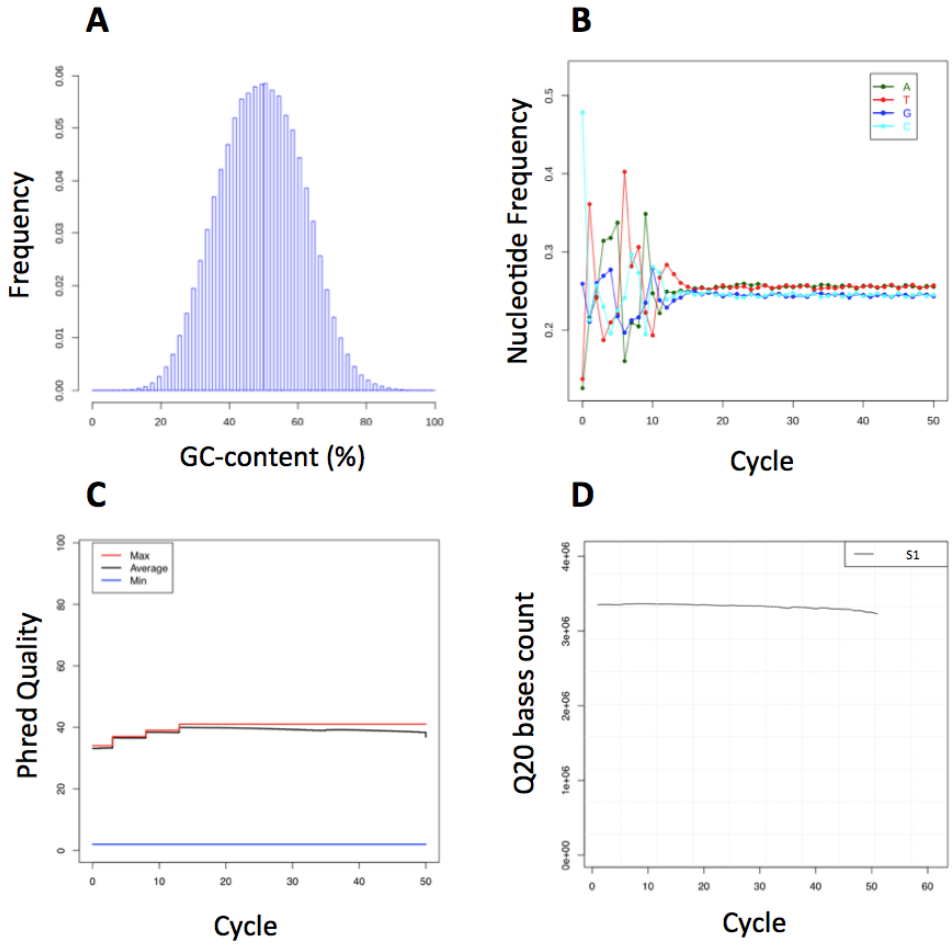

Figure 3. Examples of quality controls. (A) GC-content. PCR amplification and sequencing can be biased in GC-content. The example is showing an unbiased sample with the distribution of reads GC content symmetric and centered around 50%. (B) Base composition per cycle (or base position). Usually the first 10 cycles have an altered base composition due to the lower quality of the reads. (C). Quality in terms of Phred Quality Scores respect to the cycle. (D) Number of bases with Phred Quality Score ≥ 20 respect to the cycle.

A large number of software is currently available for quality controls. Some metrics that we used to assess the quality of our data were not available in public software and were implemented in CountSeq (*Methods, CountSeq: a flexible tool for RNAseq data analysis*).

Expression level quantification is based on the count of the number of reads, (*read count*) which map to the transcripts to be quantified. These *read counts* are

good approximations of the transcripts expression levels[33].

In order to facilitate the comparison between different genes and samples, *read counts* are usually normalized. One of the most common normalization is the RPKM (Reads per Kilobase per Million Mapped Reads)[33]. It takes into account that longer transcripts are more likely to be sequenced (resulting in a higher number of counts) and that different samples are usually sequenced with a different depth (Formula 1).

$$RPKM_{t,i} = \frac{RC_{t,i}}{L_t \times RC_i} \times 10^3 \times 10^6 \qquad (1)$$

The RPKM of a transcript *t* in a sample *i*, is defined as the *read counts* of the transcript in the samples *i* ($RC_{t,i}$) divided by the transcript length (L_g) and the total read number of the sample *i* (RC_i). Fragments per Kilobase per Million Mapped Reads (FPKM) are used when *read-pairs* are considered[34] insead of single *reads*.

Later, alternative methods based the Negative Binomial[35] distribution have been proposed to compare expression levels between samples. In particular, the edgeR[36] and DESeq[35] sofware began popular for a better estimation of the library size and for taking into account the so-called overdispertion problem. They are usually used for differentially expression analysis and recently they had been incorporated in the eQTL data analysis pipelines[37].

Different approaches have been adopted[30] for isoform-level quantification, but still there is not an unifyed consesus regarding the method to use. Based on the fact that the ENCODE project[38] and a recent large-scale eQTL[3] study used Flux-Capacitor[15] for isoform quantification, we adopted the same strategy.

RNA-seq advantages respect to other high-throughput technologies, like microarrays, are well documented in literature[39]. Here we remember that RNA-seq is not limited in detecting only transcripts corresponding to known genomic

sequence, but can identify new genes and isoforms. Furthermore, it define genes boundaries at the resolution of one base, it has a broader dynamic range of expression levels, can identify sequence variation and measure allelic imbalance.

1.5 PolyA(+) selection and rRNA depletion: different landscapes of the transcriptome

Most of the cellular RNA (60-90%) consists of structural RNAs (tRNAs and tRNAs). These transcripts are not usually the research focus so they are removed during the first step of the library construction (*Methods, RNA-seq is the most advanced technology for studying RNA samples*). Two strategies are used to reduce the amount of structural RNAs.

The first method selects polyadenylated RNAs (PolyA(+) selection). With this preparation high percentage of protein-coding mRNAs are retained and this allows to obtain after sequencing a sufficient depth to determine transcripts expression levels for most of the protein-coding genes[33] . In addition PolyA(+) selection recovers at least a subset of non-coding RNAs (ncRNA)[40].

The second method is based on rRNAs depletion (ribo-depletion). It is less common but it can potentially survey a larger set of RNA species because it detects also ncRNAs and other kind of small RNAs lacking the polyA tail. Furthermore, recent studies had suggested that ribo-depleted RNA-Seq protocols produce reliable gene expression data and are highly reproducible[41-43].

Even though many eQTL associated with the expression level changes of long non-coding RNA (lncRNA) has been reported in literature[3,4,15], there is only one lncQTL study based on ribo-depleted libraries, performed using microarrays[44]. This implies that an experiment based on the RNA-seq technology, that compares PolyA(+) and ribo-depleted samples in the context of an eQTL study is still missing in literature.

1.6 Experimental design and dataset

The experimental design is illustrated in figure 4 and consists of two parts, a pilot project (dark grey) and a main project (light-gray). This thesis is focused mainly on the pilot project, with the aim to establish which RNA enrichment approach, between ribo-depletion and PolyA(+) selection, would be better in the context of a first large eQTL study in the Sardinian population.

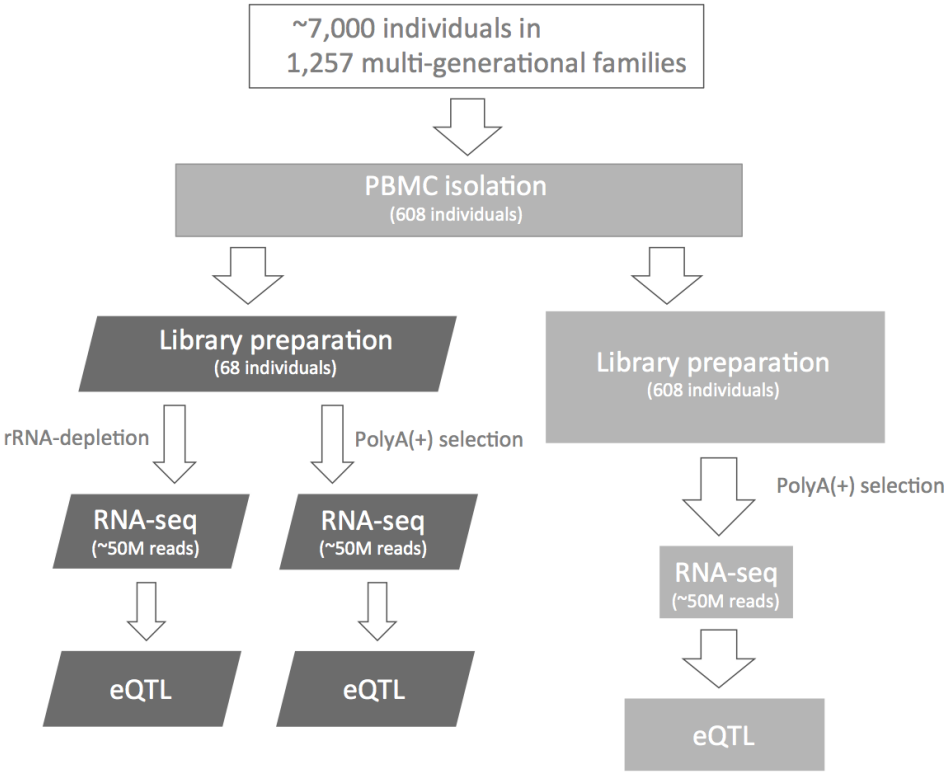

Figure 4. Experimental design workflow.

We selected a subset of 68 individuals, in order to reach the same sample size of two already published large-scale transcriptome studies based on the RNA-seq technology[15,16].

All the individuals were enrolled within the SardiNIA project[5].

The cohort consists in related and unrelated individuals, as illustrated in Figure 5.

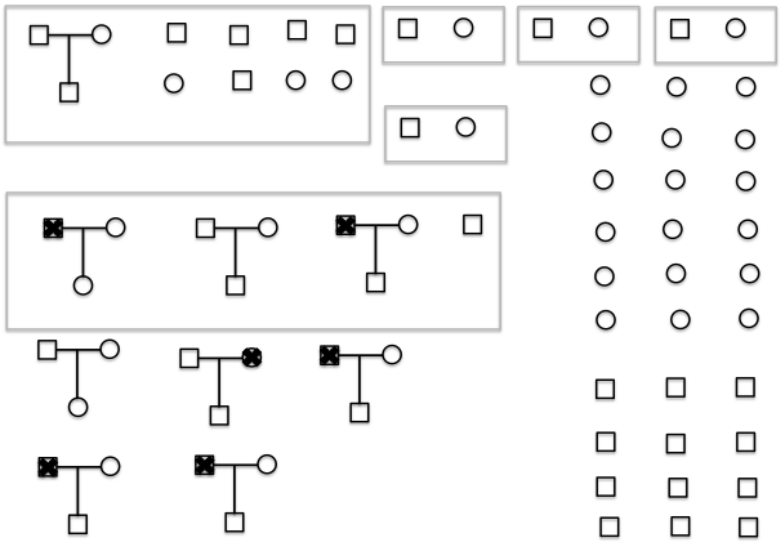

Figure 5. Pedigree of the sample set. Squares represent males and circles represent females. Individuals in the same green box belong to the same family. For the individuals with black cross, RNA sequences were not available.

As target tissue for out experiment, we decided to use the peripheral blood mononuclear cells (PBMCs). A PBMC is any blood cell having a round nucleus and consists mainly of monocytes, T-cells, B-cells, and NK cells and dendritic cells of both myeloid and plasmacytoid origin. These cells are implicated in some of the phenotypes that have been extensively characterized in the SardiNIA project[5] (like blood test results) and in autoimmune diseases that are studied in our laboratory like Multiple Sclerosis (MS)[45]. In addition, the individuals we used in our experiment were phenotyped for about 300 immuno traits[10]. Furthermore, peripheral blood is relatively easy to obtain, can be extracted with simple protocols and offers the possibility of sampling any individual at any time.

Total RNA was extracted from PBMCs and then fractioned in parallel with PolyA(+) selection and rRNA-depletion. Samples were then sequenced with

Illumina HiSeq 2000 platform. A mean of 55 millions of reads (51 nucleotides long and in paired-end) per sample were generated. This depth is comparable and even larger than the recently published eQTL studies based on the RNA-seq[3] technology.

Genotype data for the same individuals were already available. An integrated map has been generated combining data derived from four Illumina Beadchip arrays (OmniExpress, Cardio-MetaboChip, ImmunoChip and ExomeChip) and low pass whole genome sequencing performed with Illumina NGS technology. The final result was a dense map of around 17 millions SNPs.

Gene expression levels and transcripts and splice sites ratios were quantified, and genetic polymorphisms associated with expression levels changes (eQTL) were mapped.

2 Materials and Methods

2.1 PBMC isolation and RNA extraction

Blood samples were collected in EDTA vacutainer blood collection tube and immediately processed with LeukoLOCK™ Fractionation & Stabilization Kit (Ambion #AM1933) which employs filter-based leukocyte-depletion technology to isolate leukocytes from whole blood. Lysates were then obtained by flushing the LeukoLOCK Filter with 4 ml of TRI Reagent®(Ambion#AM9738). Phase separation was obtained after the addition of 800 ul of BCP (Sigma-Aldrich#B9673) followed by centrifugation. RNA was further purified and concentrated from aqueous phase, using columns contained in the PureLink® RNA Mini Kit (Ambion #12183018A). To determine the quantity and the integrity/purity of RNA samples , check controls were first performed by agarose gel (1%) electrophoresis and then by the Agilent Technologies 2100 bioanalyzer using RNA 6000 LabChip ® kit (Agilent#5067-1511) . The bioanalyzer is a bio-analytical device based on a combination of microfluidic chips, voltage-induced size separation in gel filled channels and laser-induced fluorescence (LIF) detection on a miniaturized scale. The RNA Integrity Number (RIN) software algorithm allows the classification of total RNA, based on a numbering system from 1 to 10, (with 1 being the most degraded and 10 being the most intact). All the samples with a RIN below 7,5 were discarded while the others were processed for libraries preparation and sequencing.

2.2 Ribosomal RNA depletion

Total RNA samples (4 ug each sample) were hybridized with eukaryote rRNA sequence-specific 5'-biotin labeled oligonucleotide probes provided in the

RiboMinus™ Eukaryote Kit for RNA-Seq (Invitrogen#A10837-08), which allows a selective depletion of abundant large ribosomal RNA molecules from total RNA. The rRNA/5'-biotin labeled probe complex was removed from the sample using streptavidin-coated magnetic beads (RiboMinus™ Magnetic Beads). The purified RNA samples were then concentrated using RiboMinus™ Concentration Module (Invitrogen#K1550-05).

Depletion efficiency was assessed by Agilent Bioanalyzer RNA 6000 Nano Chip and RNA concentration was defined using NanoDrop 1000 Spectrophotometer (ThermoScientific). 400 ng of purified RNA was converted into double stranded cDNA library using TruSeq RNA Sample Preparation kit (Illumina# FC-122-1001).

2.3 PolyA(+) RNA selection

PolyA(+) RNA was isolated from 4 ug of total RNA using poly-T oligo-attached magnetic beads using two rounds of purification (positive selection) as suggested in the TruSeq RNA Sample Preparation manual (Illumina # 15015050).

2.4 Library preparation

Purified samples were processed using TruSeq RNA-Seq Library Preparation Kit. Shortly, chemical fragmentation was carried out using divalent cations under elevated temperature in Illumina proprietary fragmentation buffer. First strand cDNA was synthesized using random oligonucleotides and SuperScript II (Invitrogen# 18064-014). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. After Agencourt AMPure XP beads purification (Beckman#A63882) which allows size selection of fragments, the overhangs were converted into blunt ends via exonuclease/polymerase

activities,then enzymes were removed. DNA fragments were adenylated in their of 3' ends, then Illumina TruSeq PE adapter indexed oligonucleotides were ligated, double purified and selectively enriched using Illumina PCR primer cocktail in a PCR reaction. PCR library products were purified with AMPure XP beads and quantified using the Agilent DNA 1000 assay (Agilent#5067-1504) on a Agilent Technologies 2100 bioanalyzer. The indexed individual libraries were pooled to obtain equimolar concentrations for each sample, and then processed for cluster generation.

2.5 Sequencing on HiSeq2000

Pooled libraries were loaded on a Paired End Flow Cell using the cBot System (Illumina) and the TruSeq PE Cluster Generation kit v3 (Illumina# PE-401-3001). The TruSeq technology supports massively parallel sequencing using a proprietary reversible terminator-based method that enables detection of single bases as they are incorporated into growing DNA strands. At the end of the run, 51 bp paired-end reads were generated on a HiSeq2000 instrument (Illumina) using TruSeq SBS v3 reagents (Illumina# FC-401-3001).
Finally, demultiplexed FASTQ files were generated according to the Illumina Pipeline data analysis.

2.6 Alignment and quality controls

Data analysis workflow is showed in Figure 6. Since pooled libraries were distributed in multiple flow cell lanes, the FASTQ files obtained for the same sample and deriving from different lanes were merged in a unique FASTQ file.

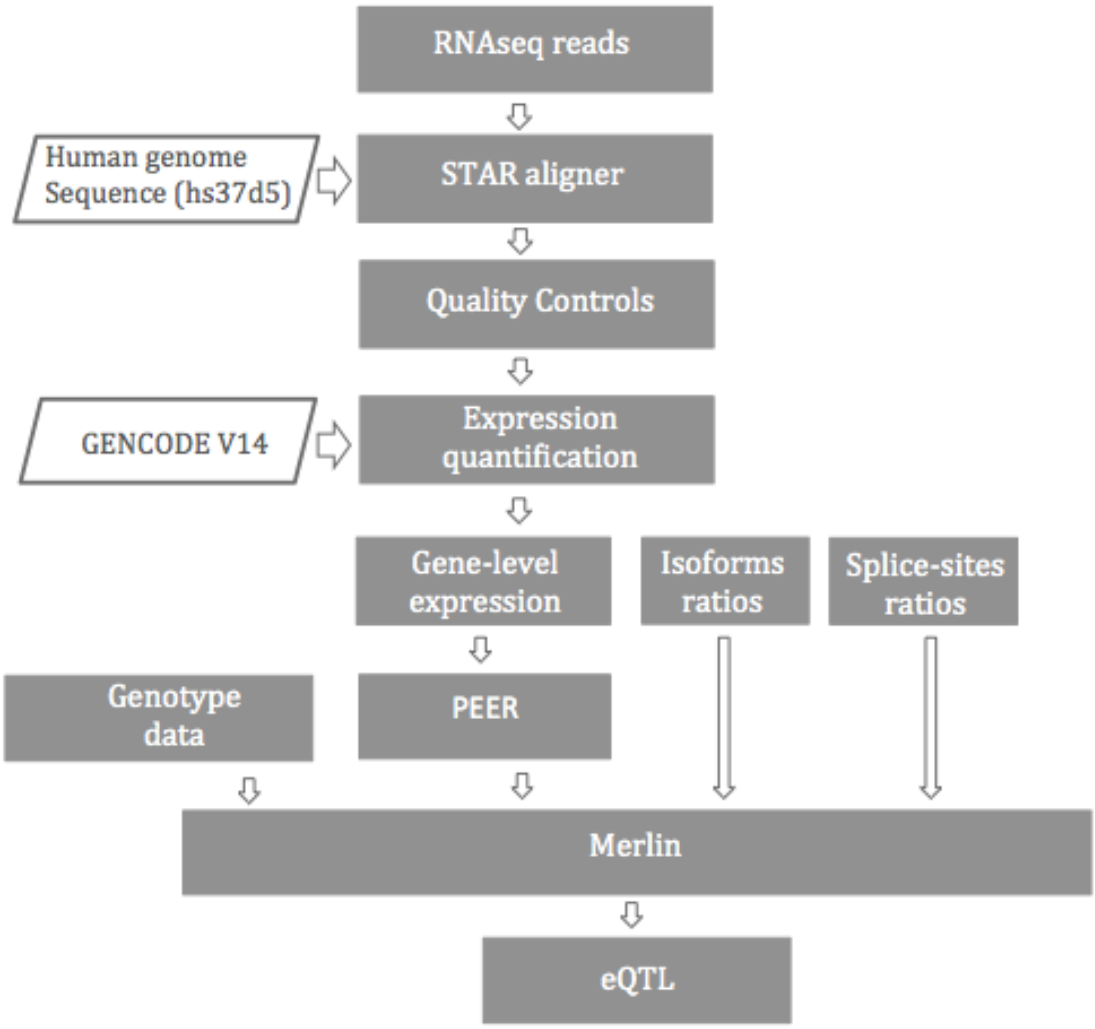

Figure 6. Data analysis workflow

Read-pairs insert-size was established by mapping the first 5 millions of reads of the FASTQ file to the ENSEMBL transcriptome, with bwa-0.5.9[46], marking duplicated and computing the insert size metrics with PICARD-1.57 (http://picard.sourceforge.net/index.shtml). All the reads were then aligned to the reference genome (h37d5) (http://www.1000genomes.org) with STAR-2.2.0c[47], spike-ins sequence (ERCC) and exon-exon junction database were generated using the GENCODE.v14 gene annotation. Quality controls were performed using SAMtools (samtools-0.1.18)[48], EVER-seq(everseq-1.0.4, http://code.google.com/p/ever-seq/), RSeQC[49] and CountSeq.

2.7 Gene-level expression quantification and hidden factor correction

For each gene we counted the number of *fragments*, or *read-pairs*, mapping to the gene (GENCODE v14 and RNAs from Repeat Masker, 60,000 genes) using the following strategy. First, we generated a custom gene-level reference collapsing the transcripts coordinates from the various isoforms described in order to obtain one annotation per gene. Overlapping regions between genes were removed. Second, fragments were counted only whether both pairs mapped to a unique locus and whether both pairs map entirely into the gene annotation (Figure 7). Fragments with only one of the read-pair mapped were counted as well. This algorithm was implemented in CountSeq.

Fragment counts were processed with DESeq[35] in order to estimate library size and stabilize the variance. Hidden factors were estimated using PEER[37] and residuals obtained and used as quantitative traits in the association analysis.

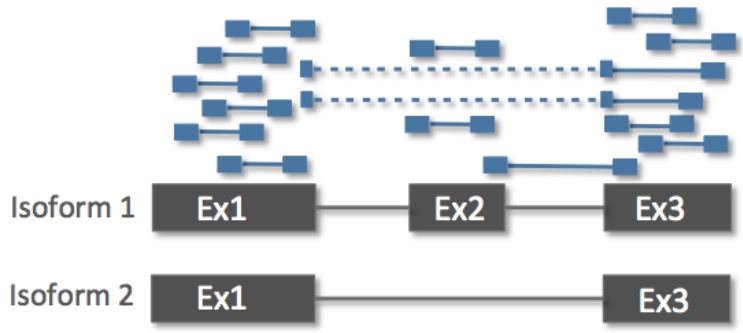

Figure 7. Reads are indicated by blue lines and boxes. Dashed lined are indicating a splicing reads.

Genes were considered expressed if FPKM mean > 0.3 and if at least the 50% of the individuals had FPKM >0. Among protein coding genes 11,114 were found expressed in PolyA(+) and 9,955 in Ribominus™ samples. Among lncRNAs 1,040 were found expressed in PolyA(+) samples and 1,023 in Ribominus™

samples.

We were interested in eQTL mapping, hence to individuate variants that are associated with changes in the expression level of the genes. Gene expression estimates can be affected by several confounding sources of variation, that can be due to experimental conditions as samples preparation differences and different operators, or biological reasons, as age, state of the cell and cell counts. Even if some of these factors can be measured (*known factors*, like age, cell counts, samples preparations) it is not possible to be aware of all potential sources of variation, hence some of them are not measured (*hidden factors*). Modeling confounding factors to correctly estimate their contribution is still a challenge and there are still open questions about how to account for them. When measured, the correct estimation of the expression levels of the additional variations due to confounders, allows a more sensitive analysis. Moreover it has been demonstrated that results in a increased number of detected associations[37]. In order to individuate and correct for confounding factors we used PEER[37] that consists of a collection of Bayesian approaches to infer hidden determinants in gene expression profiles with factor analysis methods. PEER takes as input transcript profiles and outputs hidden factors estimates. Even if PEER has been successfully used to increase the number of eQTL discoveries in several studies, the number of hidden factor should be controlled to avoid over-correction and the consequent reduction of discoveries.

In order to optimize the number of hidden factor to be removed, we used the following approach. We randomly selected 5% of the expressed protein-coding genes (560 genes for PolyA(+) and 500 for Ribominus™). We performed the association of DNA polymorphisms with gene expression levels using Merlin (Materials and Methods, eQTLs mapping) initially without applying any correction for hidden factors. Then we repeated the association analysis using five different levels of correction for hidden factors: 5, 7, 10, 15, and 17. As

shown in Figure 8, the number of discoveries increased with 5 factors and reached saturation.

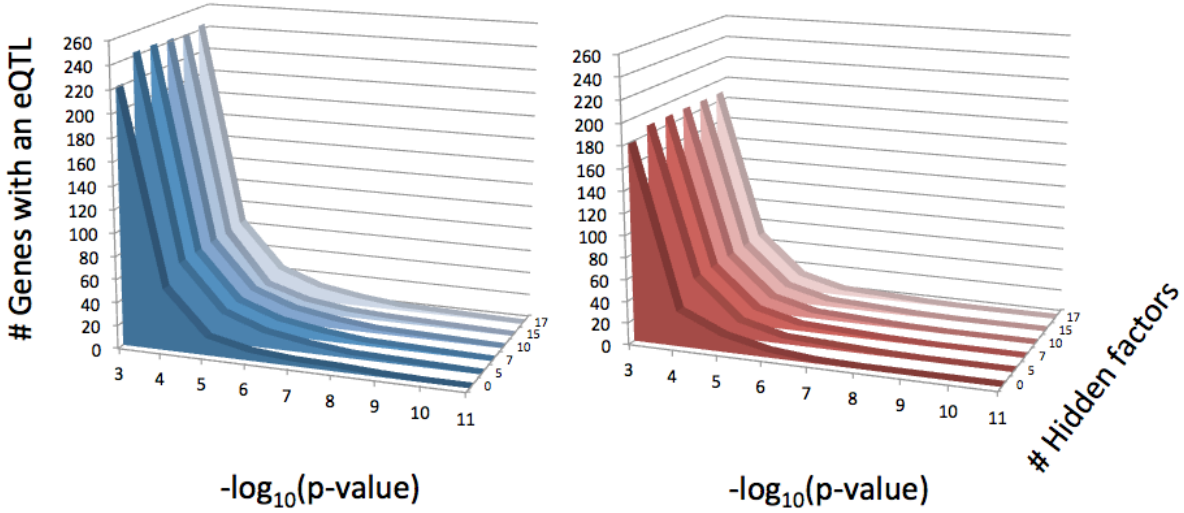

Figure 8. Number of eQTL and hidden factors

Since the correction with more than five factors didn't determine an improvement in the number of discoveries, we performed the final analysis using residuals from five hidden factors. The same number of factors was used to correct lncRNAs expression levels.

2.8 Isoform quantification

Isoforms quantification were computed by Flux-Capacitor[15] and GENCODE V14 annotation. Isoform proportions were then computed as the ratio of the RPKM of the isoform considered divided by the sum of the RPKM of all the isoforms of the corresponding gene. Isoform ratios were considered as a trait in the association analysis.

2.9 Splice-site ratios quantification

We used CountSeq to determine the number of read-pairs mapping to an exon-exon junction. We then computed two ratios for each splice-site, for (i) the exon-exon junction sharing the same donor and (ii) the exon-exon junction sharing the same acceptor. We considered all the possible combinations of exons annotated in GENCODE V14.

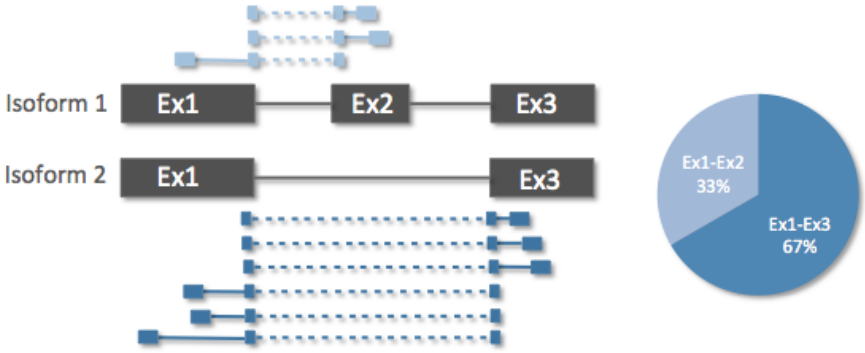

Figure 9. Read-pairs supporting a splice-site. Reads are indicated by blue lines and boxes. Dashed lined are indicating a splicing reads. Relative proportions of reads supporting different isoforms are used to quantify the splice-site ratio indicated in the pie plot.

Were considered only the combinations of exons in the same gene. In Figure 9 the read-pairs supporting a splice-site ratio site are shown. Splice-site ratios were considered as a quantitative trait in the association analysis.

2.10 eQTL mapping

We mapped eQTL using the genotype data generated by the SardiNIA project, an integrated map with ~17 millions of SNPs (*Introduction*). We filtered this variants to select only SNPs with MAF > 0.05 (based on 68 individuals) and Rsq > 0.3, obtaining 5,741,263 SNPs. The association analysis was performed with Merlin (merlin-offline)[50,51]. Before running the association analysis the expression values were standardized (we selected the –inversenormal parameter). Each gene was tested only for the associations in *cis* mapping within ± 1Mb from the transcriptional start site (TSS). Merlin has been implemented with a computationally efficient algorithm for GWAS studies, in which associations with few traits are usually performed. eQTL studies deal with tens of thousands of genes and the running time with the standard version of Merlin would be too large. Since we were interested to speed-up the analysis, we modified Merlin in order to (i) take in input an additional file of coordinates (MAP file) containing chromosome names and positions of the TSS (ii) skip the association whether the SNP-TSS distance was greater than an arbitrary value supplied by the user (iii) write in the output file genes TSS coordinates and ID, and (iv) avoid writing the association results whether the p-value was greater than an arbitrary value supplied by the user.

2.11 FDR estimation by permutation

We performed the estimation of the False Discovery Rate[52] (FDR) by permutations. Due to the large computational time required, we didn't generated an empirical *p*-value for all the tested genes but we used a median empirical *p*-value obtained from 500 randomly sampled genes. The algorithm adopted was the following.

We selected 500 genes and randomized the expression data by shuffling 1000 times the labels of the individuals. For each permutation we calculated SNPs-gene association and kept the best *p*-value (top SNP for each gene). Based on the distribution of the 1000 permuted p-values we selected an empirical *p*-value given an arbitrary threshold α (e.g. α = 0.01, first percentile of the distribution). For the 500 empirical p-values generated, we considered their median as the p-value threshold.

From the observed data (*m*, the original associations without permutations) we computed how many (*k*) had a p-value lower then the p-value threshold.

We calculated the FDR for a given permutation threshold as

$$FDR = \frac{\alpha\, m}{k}$$

We selected different values of α until the desired FDR was reached (e.g. FDR=0.05).

2.12 COViewer: An integrated viewer for multi-sample NGS data

Data visualization in RNA-seq plays an essential role. RNA-seq technology offers the possibility to detect new exons and gene boundaries and the visualization of the data is helpful in order to distinguish artifacts. One of the most used and intuitive plot is the "coverage plot", that is a scatterplot of the reads coverage against the genomic position. Different published tools for visualization of NGS data are now available[53,54] but they are not able to display (i) multiple samples on the same track, (ii) derive statistics for multiple samples based on queries provided by the user. We developed COViewer (COVerage data Viewer) in order to overcome these issues.

COViewer is implemented in C, R and Perl programming languages. It doesn't require hard pre-processing of the data. It handles BAM, and VCF.GZ and BED.GZ files indexed with TABIX[55].

COViewer finds applications also in eQTL studies, as reported in Figure 10.

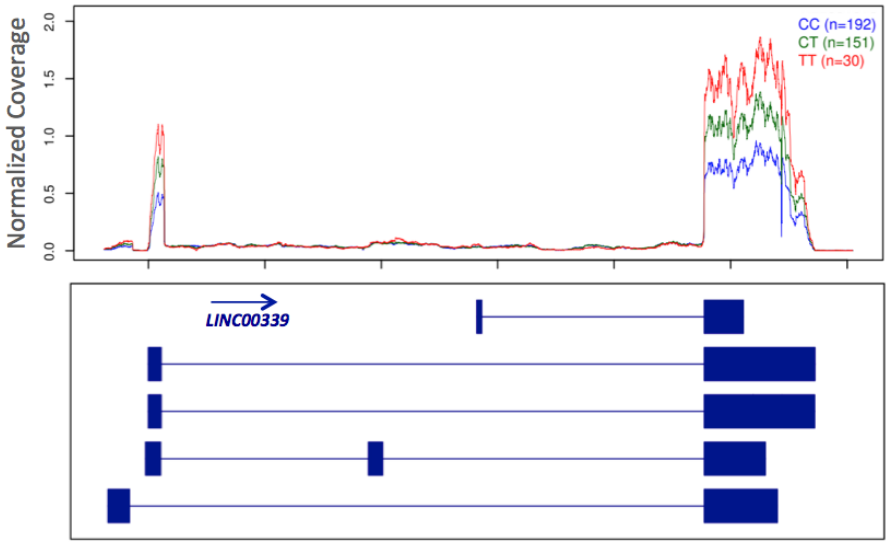

Figure 10. Example of an eQTL effect visualized with COViewer using a public available dataset[3,56]. The expression level of the LINC00339 gene is represented as coverage normalized respect to the library size (total number of reads produced for each sample). Colored lines (blue, green, red) indicate the mean expression level across the individuals for the respective genotypes (CC, CT and TT) of the genetic variant rs12061255.

The coverage plot was generated with a public available dataset[3], where rs12061255 is reported to be an eQTL for the gene LINC00339 (*p*-value=1.6e-68). The genetic variant is located 1,134 bp upstream to the transcriptional start site of LINC00339 (not shown in the plot). The expression level of the gene is represented as number of reads (coverage) normalized respect to the library size (total number of reads produced for each sample). Colored lines (blue, green, red) indicate the mean expression level across the individuals for the respective genotypes (CC, CT and TT) for the genetic variant rs12061255. As illustrated, the alternative allele T is associated with the increment of the mean expression level of the gene LINC00339.

2.13 CountSeq: a flexible tool for RNAseq data analysis

One of the main features of the RNA-seq is that a single experiment produces a huge amount of information but at the cost of dealing with high data volumes. Different tools for quality controls and quantification on RNA-seq data already exist[49,57], and perform common tasks but some of their modules are not enough flexible or are not able to analyze data directly from a remote server. We have developed a toolkit (CountSeq) that performs common tasks in flexible and efficient manner allowing remote analysis at the same time.

The CountSeq suite contains a set of tools to perform several tasks including quality controls, quantification of expression values and data visualization that can be easily combined as modules in a flexible pipeline. CountSeq is written in C language and is implemented with SAMtools[48] and Kent[56] libraries.

It deals with alignments files in BAM format[48], recognizes multiple, unique and spliced mapped reads annotation of several common aligners.

Quality Controls: There is a module for quality controls that computes several

metrics. Here we remember the (i) number of base pairs per cycle, (ii) base quality per cycle, (iii) base composition per cycle, (iv) GC-content per read, (v) number of unique reads, multiple and splicing mapping.

Gene and exons annotation: there is a module to customize the gene annotation (supplied in Gene Transfer Format, GTF), in order to facilitate the quantification of expression levels. It can collapse the transcript isoforms in a single gene reference and re-annotate introns, 5'UTRs and 3'UTRs.

Quantification of gene expression: a module for the quantification of gene expression is available. Given an annotation file in GTF format, it computes genes, exons, intron and splice sites reads (and read-pairs) counts.

Visualization: the *mpileup* module can be easily used to generate wiggle (WIG) and bedGraph files, which are commonly common format to represent a signal track of mapped reads. Furthermore starting from a BAM file CountSeq can generate a bed file with exon-exon junctions genomic coordinates, quality controls and coverage informations that can visualized in common genome browsers[56].

3 Results

3.1 Ribominus™ and PolyA(+) libraries are largely different in RNA classes composition.

Ribominus™ and PolyA(+) samples were sequenced similar mean depths, respectively of ~56.45 (± 11.70) and 54.27 (±6.94) millions of read-pairs (Figure 12).

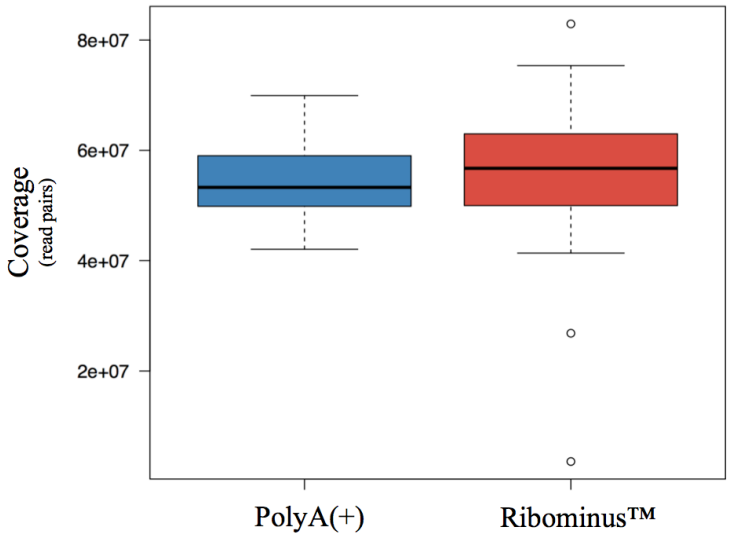

Figure 12. Ribominus™ and PolyA(+) coverage reported as millions of read pairs.

The coverage was variable between individuals and this behavior was present in both libraries: in Ribominus™ preparation the standard deviation (11.70) of the coverage was larger than in PolyA(+) libraries (6.94) and this was due to the very low depth obtained in two samples (Figure 12). Coverage variability between samples had been reported in literature and it seems to be a common bias of the RNA-seq technology. Still there is not a unified consensus on how can be obtained the reduction of this variability. We tried to understand the source of this bias taking in consideration different variables (number of PCR

cycles, operator, RNA quality, and reagents stocks) but we didn't find any reasonable explanation.

The software used here for the expression quantification deals with reads that map to unique genomic loci (*Methods*). Furthermore, the software used for splicing and isoform quantification made a large use of reads spanning exon-exon junctions. For these reasons we were interested in the composition in both libraries of reads classified as mapping to unique or multiple loci, and to splicing or not splicing sites. Statistics are reported in Figure 13.

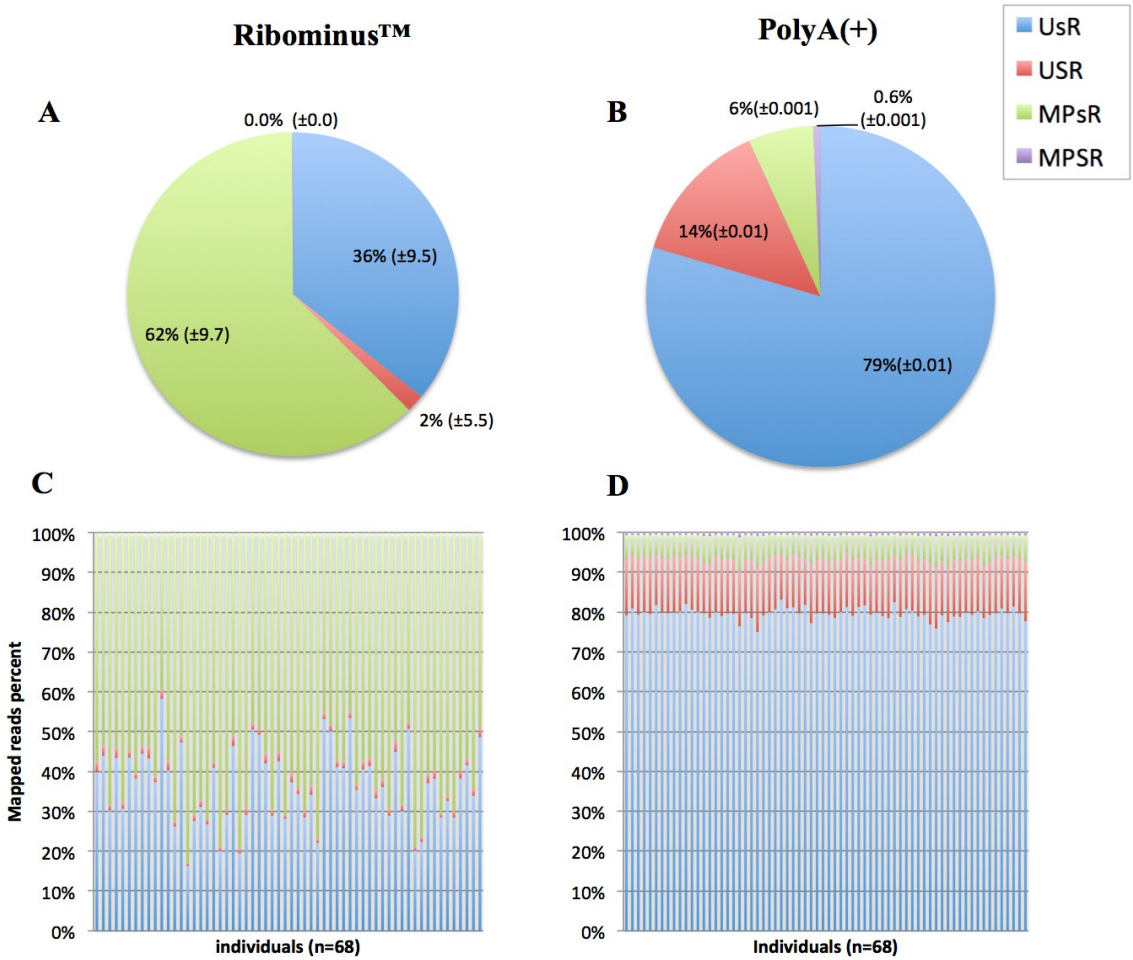

Figure 13. Ribominus™ and PolyA(+) reads composition. (A-B) Mean (±SD) among 68 individuals. (C-D) Composition, reported as percentages, by individuals. Multiple mapped spliced reads (MPSR), multiple mapped not spliced reads (MPsR), uniquely mapped spliced reads (USR), and uniquely mapped not spliced reads (UsR).

The composition of read classes was largely different between libraries. PolyA(+) samples were enriched in uniquely mapped reads (UsR) and in spliced uniquely mapped reads (USR), while Ribominus™ samples were dominated by multiple mapped reads (MPsR).

In order to define the source of the differences in read composition, we re-annotated each read class according to which location they mapped respect to the annotated gene (exonic, intronic and intergenic regions respectively). Results are shown in Figure 14.

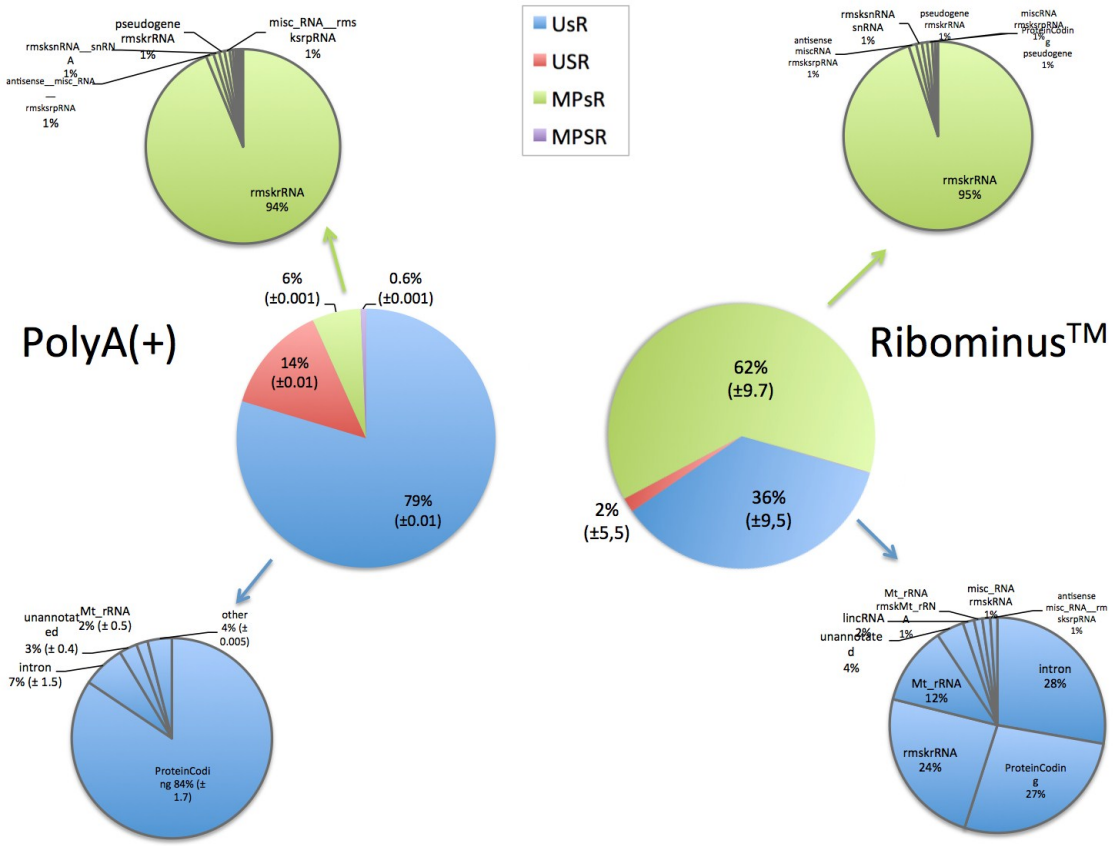

Figure 14. Read composition segmented by gene category in Ribominus™ and PolyA(+) samples. Means (±SD) among 68 individuals are shown. Multiple mapped spliced reads (MPSR), multiple mapped not spliced reads (MPsR), uniquely mapped spliced reads (USR), and uniquely mapped not spliced reads (UsR). Genes from repeated masker are indicated with rmsk followed bythe standard name (for example, rmskrRNA stands for rRNA from Reapeat Masker database).

Multiple mapping reads (MPsR) were mostly coming from rRNAs as annotated in RepeatMasker database (rmskrRNAs) (94% in PolyA(+) and 95% in Ribominus™ samples) (Figure 14). Since multiple mapped reads were the vast majority of the Ribominus™ preparations, rRNA depletion in these libraries was not efficient. We hypothesized that this was mostly due to the low number of rounds of depletion we performed. As described in *Methods*, we ran only one cycle of depletion and this approach we obtained samples with still 76% of the total amount of mapped reads coming from rRNA transcripts (~63% represented by multiple mapped reads and ~13% by unique mapped reads).

Unique mapped reads (UsR) showed a complex behavior. In PolyA(+) samples, 84% of them derived from protein-coding exons (Figure 14), 7% from intronic sites, 3% from inter-genic regions (un-annotated), 2% from mitochondrial rRNAs and 4% from all the remaining categories.

A very different pattern was observed in Ribominus™ samples. The most enriched categories were intronic sites (28%), protein-coding exons (27%), genomic rRNAs (24%) and mitochondrial rRNAs (14%) (Figure 14).

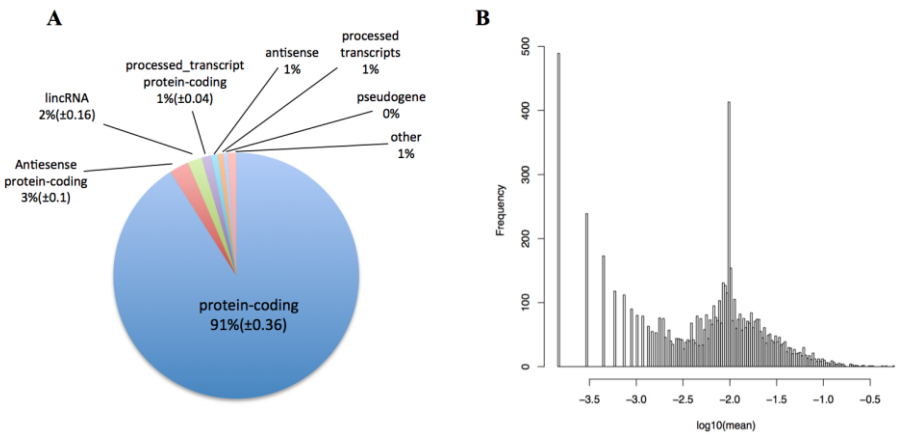

Figure 15. Unique mapped intronic reads composition in Ribominus™ libraries. (A) Composition by gene types. Means among 68% (±SD) individuals are shown. (B) Composition by single gene. Overlapping introns sites between different genes are considerated as separate signals.

Intronic reads came mostly from protein-coding introns (Figure 15A) and were broadly distributed among genes (Figure 15B): we observed a peak for moderate contribution a 0.01% ($\log_{10}(0.01)=-2$) and for absent expression at 0.0% (coming from those genes that are not expressed).

The first four genes that contributed to the whole intronic fraction in Ribominus™ samples were PTPRC (0.57%) (Figure 16A), NAMPT (0.50%) (Figure 16B), MBNL1 (0.42%) (Figure 16C), and FAM65B (0.34%) (Figure 16D).

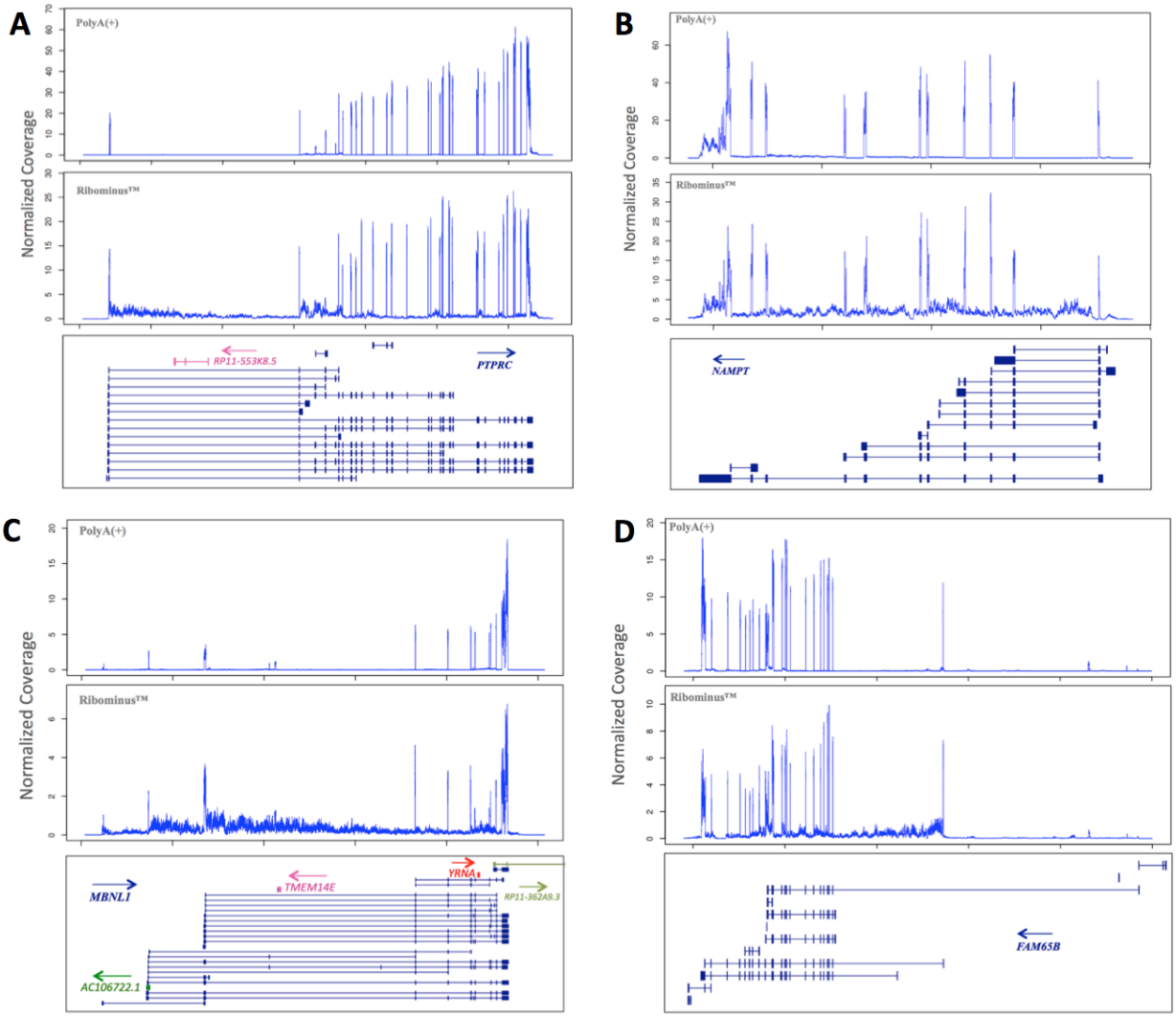

Figure 16. Coverage plots for the first four genes that contributed to the whole intronic fraction in Ribominus™. (A) PTPRC, (B) NAMPT, (C) MBNL1 and (D) FAM65B. The expression level of the gene is represented as number of reads (coverage) normalized respect to the library size (total number of reads produced for each sample) (see in the introduction the section *Coviewer: An integrated viewer for multi-sample NGS data* of a for more details about how coverage plots were generated).

Intronic reads were spread out along the introns (Figure 16) supporting the hypothesis that they don't represent the processed mature RNA molecules but the immature transcripts. Immature transcripts shouldn't be polyadenylated and this is in line with the fact that we observe them mostly on Ribominus™ samples. Large fractions of intronic reads have already been observed in total RNA preparations[58], and it has been hypothesized that these reads characterize mostly genes subjected to intense alternative splicing regulation.
From the coverage plot in Figure 16, and in particular for PTPRC (Figure 16A) and MBNL1 (Figure 16C), some introns presented stronger signal than those of the respective exons. Since signals from low expressed exons can be entirely hidden by intronic signals (Figure 16) we hypothesized that the isoform assembling and quantification in Ribominus™ samples could be less accurate than PolyA(+) libraries, particularly for low expressed isoforms. We hypothesized that, while the proportion of rRNA could be reduced with a higher number of depletion cycles, the large intronic reads fraction seems to be an intrinsic feature of the Ribominus™ preparation.

Unique mapped splicing reads (USR) were a considerable portion in PolyA(+) samples and most of them came from protein-coding exon-exon junctions (97% in PolyA(+) and 95% in Ribominus™ preparations) (data not shown).

Based on these results we estimated that the number of reads available for expression quantification were 93% in PolyA(+) and 38% in Ribominus™ samples respectively. This implies that, if we would start from a fixed amount of reads, like 50 millions (M), we will obtain 45M of useful reads in PolyA(+) and 19M of reads in Ribominus™ libraries. Considering only protein-coding genes, the difference would be even larger, in fact we will have 40M reads in PolyA(+) and 6M in Ribominus™ preparation respectively. Based on all these

observations, it is expected that both gene-level expression, splicing and isoform quantification would be more sensible and accurate using PolyA(+) preparations, with an improved statistical power and an higher number of eQTL discovered.

Notably a portion of reads in both preparation were found outside the gene boundaries (un-annotated), in particular 4% in the Ribominus™ and 3% in the PolyA(+) preparation. Although a subset of these reads could represent noise signals, we cannot exclude that they could belong to un-annotated genes. We plan to perform different the *de-novo* assembling strategies to assess the source of these inter-genic reads.

3.2 The first collection of eQTL in the Sardinian population

From the RNA-seq data we determined the expression levels of the genes, isoforms and splicing sites in PBMC samples, and then we mapped the genetic polymorphisms associated with changes in the expression levels considered as quantitative traits (*Methods*). We generated several lists of genetic variants associated with the expression of protein-coding genes (eQTL), long non-coding RNAs (lincQTL), splice-sites ratios (sQTL) and isoforms ratios (isoQTL). In order to stratify the discoveries with different levels of confidence, we segmented the QTL by different values of False Discovery Rate (FDR) estimations (1%, 3% and 5% respectively) (*Methods*). The number of significant QTL is reported in Table 1 and Table 2.

Table 1. QTL in PolyA(+) samples.

trait	#traits	#tests	# traits with a QTL		
			FDR 1%	FDR 3%	FDR 5%
Protein-coding (eQTLs)	11,114	45,458,693	540	609	697
lncRNA (lncQTLs)	1,040	4,280,805	103	121	145
protein coding splicing (sQTLs)	19,755	80,238,841	986	1,115	1,226
proteincoding isoforns (isoQTLs)	99,533	401,902,931	1,460	1,650	1,814

Table 2. QTL in Ribominus^{TM} samples.

trait	#traits	#tests	# traits with a QTL		
			FDR 1%	FDR 3%	FDR 5%
Protein-coding (eQTL)	9,955	40,812,502	253	287	326
lncRNA (lncQTLs)	1,023	4,164,177	37	40	48
protein coding splicing (sQTL)	2,191	9,107,873	33	33	36
proteincoding isoforns (isoQTL)	89,972	363,555,750	1,109	1,248	1,424

The number of eQTL discovered was comparable with those already reported in other RNA-seq studies with a similar sample size. Comparing PolyA(+) selected samples (analyzed and processed in the same way as the considered published datasets), we found 540 eQTL at FDR 1% while previous studies reported 256 eQTL at FDR 43% in Caucasians[15] and 411 at 1% FDR in Africans[16]

PolyA(+) selection allowed the detection of more QTL than those resulting from Ribominus™ protocol (Table 1 and 2). In order to test if the source of this difference was due to technical reasons or to differences in the statistical power (that we hypothesized was due to the difference in read composition, see previous chapter) we determined the sharing degree of QTL between PolyA(+) and Ribominus™ libraries.

An exact intersection of the discoveries was not suitable to establish the degree of the overlap, because of the different statistical power between of the two libraries preparations (45% of the reads were suitable for the quantification in PolyA(+) samples and only the 19% in the Ribominus™ preparations). In order to perform the comparison, we used the $\pi_1=1-\pi_0$ statistic[52], that provides an estimate of the replication rate with a continuous statistic that does not depend on the selection of a significance threshold. Furthermore, it accounts for the number of tests performed. The results using the π_1 statistic are reported in Table 3. We observed a large sharing of QTL between libraries preparations. The maximum level of sharing included QTL associated with changes in protein-coding genes (eQTL in table 3) (95%).

Table 3. QTL sharing

Associationı (FDR 5%)	PolyA in Ribominus™ π_1(#tests)
eQTL	0.95 (544)
lncQTL	0.81 (94)
sQTL	0.78 (150)
isoQTL	0.0 (1576)

For the isoform expression level changes we obtained a minor overlap and this was mostly evident for the lowest expressed isoforms. In fact, as reported in Table 4, considering only associations for the most expressed transcripts the overlap increased, reaching the 71%. As reported in the first column of Table 4, we required a gene expression level with FPKM>0 in at least the 50% of the individuals, and a FPKM mean in the population of at least 0.3. The second column of Table 4 reports the expression level filters considered for the isoforms. Unlike gene expression levels, isoform expression levels are reported in RPKM because for transcript quantification we used Flux-Capacitor that considers reads instead of read-pairs (see *Methods*).

Table 4. isoQTL sharing. The number of individuals required to pass the expression level filter are indicated in parenthesis.

GeneLevel ExpressionFilter	TranscriptLevel ExpressionFilter	#isoforms	#genes	PolyA in Ribominus π_1
FPKM>0.0(50%), FPKMMean>0.3	-	1576	1237	0.00
FPKM>0.0(50%), FPKMMean>0.3	RPKM>0.0(50%)	327	254	0.55
FPKM>0.0(50%), FPKMMean>0.3	RPKM>0.0(50%), RPKMMean>0.1	279	214	0.58
FPKM>0.0(50%), FPKMMean>0.3	RPKM>0.0(50%), RPKMMean>0.3	203	154	0.59
FPKM>0.0(50%), FPKMMean>0.3	RPKM>0.0(50%), RPKMMean>1	91	69	0.71

We hypothesized that the correlation between the isoform expression level and

the number of overlapping discoveries between different libraries was mostly due to the fact that isoform quantification was less accurate in Ribominus™ samples. In particular, as discussed in the previous chapter, the reduced amount of reads suitable for the quantification and the larger number of reads belonging to intronic regions would have negatively influenced the efficiency of the isoforms reconstruction and hence the quantification.
Similar consideration we made for the isoQTL can be applied to the other eQTL. Therefore we concluded that the difference in the number of discoveries between the two preparations was not due to technical biases but to the reduced number of reads available in Ribominus™ samples.

We tested only polymorphisms in *cis* (*cis*-eQTL) respect to the target gene (± 1Mb from the transcriptional start site). *Trans*-eQTL are in general less studied because they are more difficult to map. They require higher significant thresholds (because all the regions of the genome must be tested). Since the main aim of this thesis was to evaluate differences in discoveries obtained with the two library preparations, we considered that the study of only *cis*-eQTL would be more robust to reach our objective.
eQTL usually map near the Transcriptional Start Site (TSS) of the target genes, and in particular the hypothesis is that *cis*-eQTL may have a direct role on the regulation of the nearby transcripts. In line with already published results[16], we found that eQTL are distributed around the target genes. The same behavior was observed for the sQTL, as illustrated in Figure 17. In order to evaluate potential biases, we generated a random distribution with permuted data (500 genes/splice-sites by 1000 permutations), indicated with dashed lines in Figure 17.

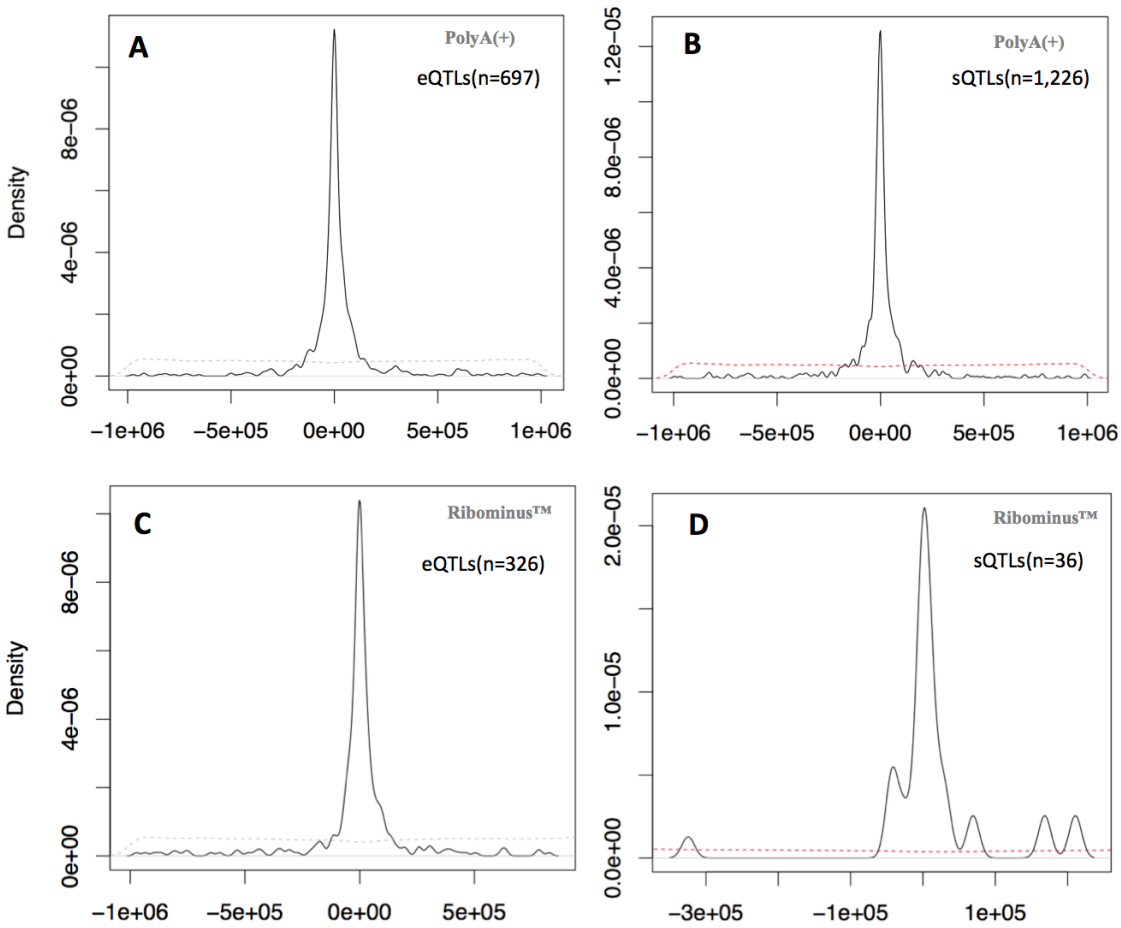

Figure 17. Distribution of the distance between topSNPs respect to the TSS of the target genes. A random behaviour is indicated with dashed lines and was calculated using the distance of the topSNPs from the TSS given by the permuted data (500 genes/splice-sites by 1000 permutations). Relative distances respect to the TSS are reported in base pairs.

The distribution around the TSS was observed in both PolyA(+) and Ribominus™ preparations, indicating that the discoveries were consistent between the two protocols and thus once again supporting that the difference in the number of QTL was due to a difference in statistical power rather that a technical bias.

3.3 COViewer and examples of sQTL

In this section we will show how COViewer, the browser we implemented, provides a powerful visual inspection of multi-sample NGS data. Visual inspection is particularly helpful in the contest of the alternative splicing characterization. In fact, alternative splicing events are still hard to study even with an advanced technology like the RNA-seq. The main reason is that NGS methods produce short-reads, hence the full-length sequence of the transcripts is not directly available and therefore has to be inferred. An additional complication is given by the fact that different splicing isoforms can share the same exon and the same exon-exon junction, and in these cases could be difficult to establish from which alternative transcript the signal originates. Up until now, large progress has been made in the contest of isoform quantification with the RNA-seq technology, but expression-level estimates still vary widely across methods[30].

All these considerations imply that alternative splicing characterization and therefore sQTL discoveries, require additional validation. Visual inspection offers a powerful way to verify alternative splicing inference, in particular coverage plots. Coverage plots are scatter plots of the expression level signal (normalized respect to the library size) and the genomic position, see *Methods*). They allow to inspect the overall signal of genes, exons and introns, and distinguish between signals and noise.

With the following examples we will show how coverage plots can be used to describe alternative splicing events in the contest of sQTL.

The variant rs2256974 (mapping to the genomic position 6:31555392) is an example of sQTL. This variant was associated with the alternative splicing of the exon 3 of the gene LST1. The effect of the association is shown in the coverage plot in Figure 18.

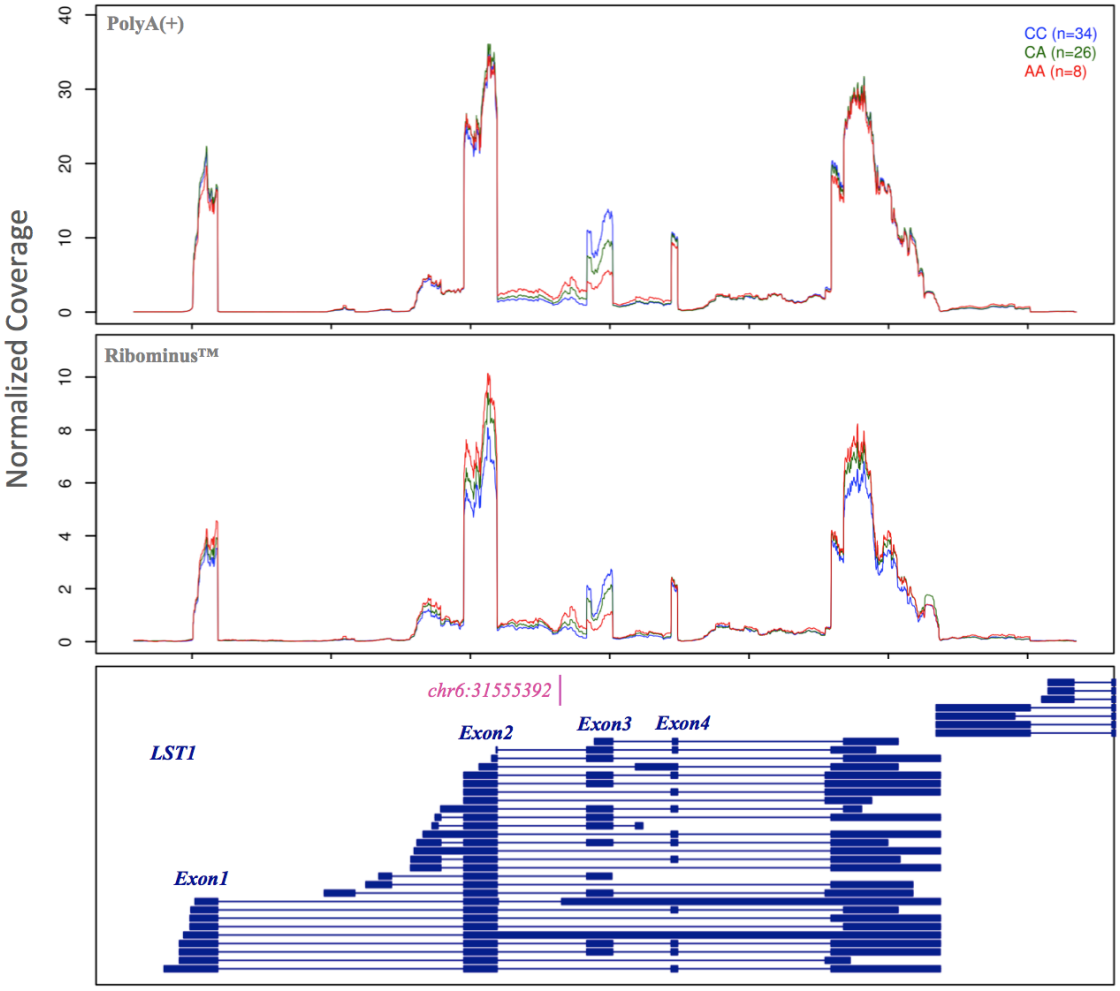

Figure 18. Coverage plot of LST1 gene, segmented by SNP rs2256974 that was associated with the alternative splicing of LST1. Genomic region 6:31553791-31557174.

The variant rs2256974 was the top SNP (SNP with the lowest p-value) in PolyA(+) samples and was associated with FDR 5% in Ribominus™ preparation. It was correlated with the decrease of the proportion of the exon-exon junction spanning exon2 and exon 3. The association effect (β) had the same strength and direction in the two library preparations (β=1.243 in Ribominus™ and β=1.219 in PolyA(+) samples) (Table 5). Furthermore, the variant was not associated with the gene-level expression of LST1 (p-value=0.4901 in PolyA(+) and p-value=0.2032 in Ribominus™ samples, Table 5), supporting the hypothesis that the SNP rs2256974 directly regulates the exon 3 skipping.

Table 5. Association statistics of SNP 6:31555392

SNP	Allele1	Allele2	Trait	Trait type	PolyA		Ribominus	
					β	p-value	β	p-value
rs2256974	C	A	6:31555095___31555418 (Ex2-Ex3)	splicing site	1.243	7.49E-13	1.219	4.66E-12
rs2256974	C	A	ENSG00000204482.5 (LST1)	gene-level	-0.125	0.4901	-0.223	0.2032

Another example of sQTL is the variant rs4844390 (mapping to the genomic position 1:207934849) that was significantly associated with the alternative splicing of the gene CD46. rs4844390 resulted the topSNP for the Ribominus™ samples and was significantly associated also in PolyA(+) preparations at FDR 5% (Table 6).

Table 6. Association statistics of SNP 1:207934849

SNP	Allele1	Allele2	Trait	Trait type	PolyA		Ribominus	
					β	p-value	β	p-value
rs4844390	A	G	1:207943666___207941168	splicing site	1.396	3.80E-11	1.412	1.79E-11
rs4844390	A	G	ENSG00000117335.13 (CD46)	gene-level	-0.392	6.23E-02	-0.497	0.01805

The coverage plot of CD46 gene, segmented by the rs4844390 genotype, is illustrated in Figure 19. The variant rs4844390 is intronic and was associated with the skipping of the exon indicated by the gray arrows in Figure 19. The global expression of the gene is weakly correlated with the genomic variant (not significant at FDR 5%) but with opposed effect respect to the skipped exon, supporting the hypothesis that the variant is directly influencing the splicing event and not the overall expression of the gene. This is indicated by the sign of the association effect (β) that is discordant between gene-level and the exon-exon junction, reported as 1:207943666___207941168 (Chromosome:IntronStop__IntronStart) in Table 6.

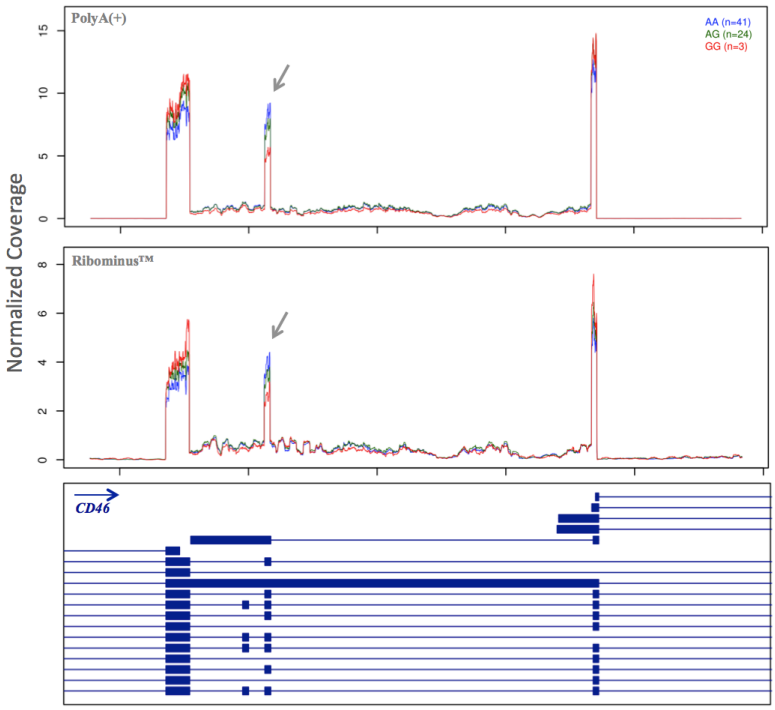

Figure 19. Coverage plot of CD46 gene, segmented by SNP rs4844390 that was associated with the alternative splicing if CD46. Genomic region 1:207940182-207943922.

3.4 Uniqueness of the genome and uniqueness of the discoveries

The extensive characterization of the DNA and of the RNA of the Sardinian population achieved by our experimental design, offered the opportunity to establish whether the uniqueness of the Sardinian genome (*Introduction*) can uncover regulatory mechanisms that cannot be studied in non Sardinians.

eQTL studies focus on common variants, usually with MAF>0.05 (Methods), but the pattern of allele frequencies varies between populations. A fraction of variants that are common in Sardinians could be rare (MAF≤ 0.05) in other populations. Hence, if any of these variants were also eQTL, we will have the unique opportunity to dissect regulatory mechanisms that can be better characterized or even uniquely studied in the Sardinians. In Table 7 and Table 8 are shown those eQTL that have respectively MAF≤0.05 and MAF≤0.01 in the 1000 Genomes catalog.

Table 7. Sardinian-private QTL (1000G MAF≤ 0,05). Not available data are reported as NA.

SNP	#SNPs (FDR5%)	#SNPs (FDR5%) called also in 1000G	Library	MAF≤ 0,05(NA)				
				1000G	ASN	AMR	AFR	EUR
eQTL	690	668	PolyA(+)	34(0)	61(47)	30(0)	93(8)	21(1)
		318	Ribominus™	19(0)	41(15)	16(0)	41(6)	12(1)
lncQTL		140	PolyA(+)	7(0)	16(6)	7(0)	21(2)	7(0)
		48	Ribominus™	1(0)	4(1)	1(0)	5(0)	1(0)
sQTL		1226	PolyA(+)	29(0)	60(37)	28(0)	83(4)	22(0)
		36	Ribominus™	0(0)	3(0)	0(0)	1(0)	0(0)

Table 8. Sardinian-private QTL (1000G MAF≤ 0,01). Not available data are reported as NA

SNP	#SNPs (FDR5%)	#SNPs (FDR5%) called also in 1000G	Library	MAF≤ 0,01(NA)				
				1000G	ASN	AMR	AFR	EUR
eQTL	668	668	PolyA(+)	7(0)	19(47)	7(0)	33(8)	4(4)
	318	318	Ribominus™	5(0)	17(15)	3(0)	14(6)	1(1)
lncQTL	140	140	PolyA(+)	2(0)	11(6)	1(0)	7(2)	2(0)
	48	48	Ribominus™	0(0)	3(1)	0(0)	2(0)	1(0)
sQTL	1226	1226	PolyA(+)	2(0)	30(37)	4(0)	36(4)	3(0)
	36	36	Ribominus™	0(0)	2(0)	0(0)	0(0)	0(0)

The p-value ranged from 6.703e-10 to 2.536e-05 (Figure 20A) and their position was centered near the transcriptional start site (TSS) of the target gene (Figure 20B).

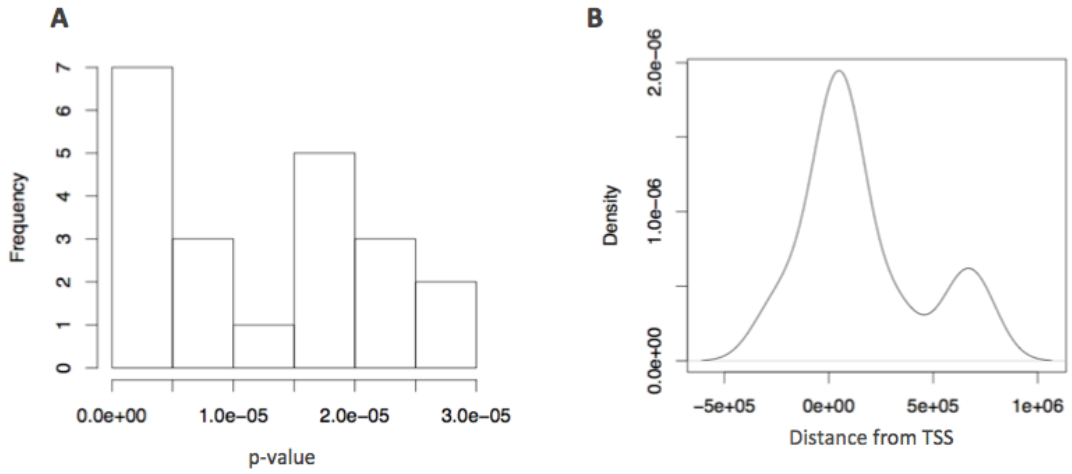

Figure 20. p-value and relative distance from TSS for Sardinian eQTL with European MAF≤0.05. Relative distance is indicated in bases pairs.

We expect to dramatically increase the number of Sardinian-private eQTLs when we will repeat the same analysis on the entire cohort (608 individuals).

3.5 eQTL reproducibility in the entire cohort

Given the larger number of discoveries using the PolyA(+) selection (Table 1 and Table 2), we applied this protocol to the whole sample set (608 indivuduals). Most of the analysis is still in progress, however, here we will show some preliminary result and the degree of the overlap of discoveries between different experiments.

The analysis strategy was the same adopted for the 68 individuals (*Introduction and Methods*).

The number of eQTL obtained from 608 individuals is illustrated in Figure 21.

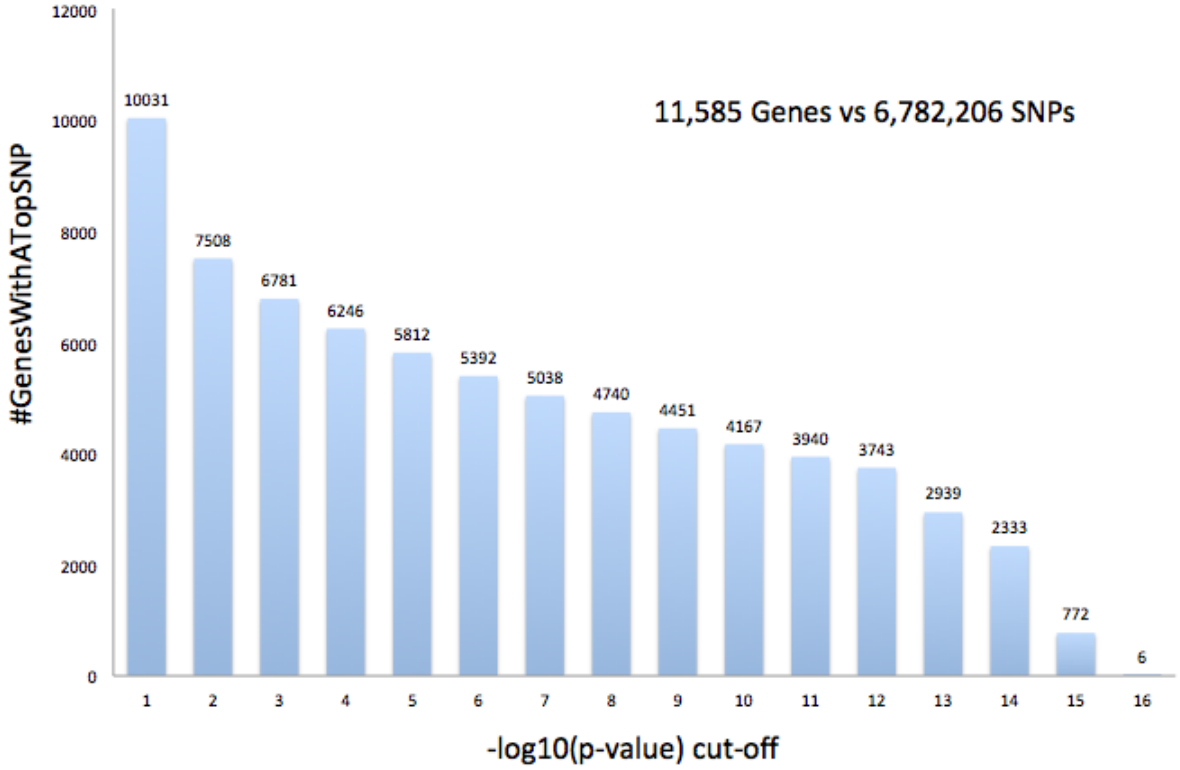

Figure 21. Number of protein-coding eQTL discovered, stratified by arbitrary p-value cut-offs

eQTL are stratified by different arbitrary p-value thresholds (Figure 21). The estimated overlap, between the eQTL lists obtained from 68 individuals and 608 individuals, was very large (π_1=0.99). In table 9 is also reported the overlap with

our results and the available public datasets from eQTL studies. Comparison are segmented by experiments and by gene category (protein-coding and lncRNAs).

Table 9. Overlap with public available datasets

AssociationType	FDR 5% (# tests)	ALL	π1 (# tests)
lincQTL	LCLs_EUR_373 (218)	PBMC_PolyA_68	0.71 (80)
		PBMC_Ribominus_68	0.79 (66)
eQTL (protein-coding)	LCLs_EUR_373 (2079)	PBMC_PolyA_68	0.48 (1228)
		PBMC_PolyA_608 (20 factors)	0.85 (1240)
		PBMC_Ribominus_68	0.36 (1078)
	Blood_EUR_922 (10326 ids in total, 680 of which with no match)	PBMC_PolyA_68	0.45 (7731)
		PBMC_PolyA_608 (20 factors)	0.92 (7984)
		PBMC_Ribominus_68	0.37 (6836)

In general, the reproducibility between our results and the public available eQTL was large and was correlated with the sample size of the compared datasets: for the pilot project, the reproducibility with the other eQTL studies ranged from 36% to 79%, while for our largest dataset (608 individuals) the overlap was estimated to be about 85% with LCLs from Europeans[4] (373 individuals) and 92% with whole blood from Europeans[3] (922 individuals) respectively.

4 Conclusions and future plans

In this work, by running a pilot study on RNA extracted from PBMCs from a relatively small sub-set of Sardinian individuals (N. 68), we demonstrated the impact of two different kind of libraries preparations, PolyA(+) selection and rRNA-depletion, in the contest of an eQTL study.

We found that even though with the RibominusTM preparation we were potentially able to quantify a broader portion of RNA types, with the PolyA(+) selection we obtained a largest number of eQTLs associated with both gene-level expression, splicing and isoform proportion changes. This difference in the number of discoveries was mostly due to the lower statistical power of the Ribominus^{TM} preparations. In fact, the Ribominus^{TM} protocol was not efficient in the depletion of rRNA transcripts, and this reduced the possibility to characterize all the other RNA species, like protein-coding and long non-coding RNAs, resulting in a lower number of discoveries. We thus concluded that in the context of a first a large-scale study in the Sardinian population, the PolyA(+) protocol would be more advantageous than the rRNA depletion.

Still, there are some interesting findings that come up from the analyses performed in this relatively small subset of samples. We studied only cis-eQTL (eQTL that map within ±1Mb from the transcriptional start site of the target gene), and in line with the previously reported studies, we found that cis-eQTL map very close to the target transcripts, suggesting once again that these variants directly regulate the expression level of their nearby target genes.

As expected by the fact that the Sardinian population is a founder population, we discovered new eQTL that are easier to detect, or perhaps could be detected only, in Sardinians, because they are rare, and hence hard to be characterized, in other populations.

Furthermore, we sequenced the RNA of a larger cohort of individuals (N.608)

using the PolyA(+) selection and performed some preliminary eQTL analysis in this larger sample set.

We found that eQTLs from published datasets showed high reproducibility in our eQTL lists, suggesting that, as previously reported[4], the overlap of eQTL obtained from different experiments, including those performed on RNAs obtained from a different source (such as cell lines) is large.

However, many issues remain to be addressed. By extending the analysis of the pilot project, we plan to better characterize the lncRNAs classes. We expect that the Ribominus^{TM} samples will allow us to study the full landscape of coding and long non-coding RNAs, hence both polyadenylated and non-polyadenylated forms. In particular, we are interested in establishing whether lncRNAs share the same eQTL with nearby protein-coding genes, in order to assess if they have an independent genetic control or not, and to evaluate potential reciprocal regulatory mechanisms with nearby genes. Thus far there is only one published study that provided insights about this link, but it was performed using microarray expression profiling[44]. We intend to inspect this issue by analyzing RNA-seq data, which in contrast with array data is not subjected to any ascertainment bias and has a much broader dynamic range; it is thus more accurate for the characterization of low expressed transcripts as the long non-coding RNAs.

Finally, many new results are expected from the completion of the analyses of the larger cohort of 608 individuals. Our samples are part of the SardiNIA project and, as such, have been extensively phenotyped for over 800 quantitative traits since 2001, when the project has begun. We anticipate that the integration of all the available information will allow us to dissect the molecular relationships among genetic variation, gene expression and complex traits at an unprecedented level of resolution.

5 Reference

1. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat Rev Genet* **10**, 184–194 (2009).
2. Dermitzakis, E. T. From gene expression to disease risk. *Nat Genet* **40**, 492–493 (2008).
3. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
4. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research* (2013).
5. Pilia, G. *et al.* Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLoS Genetics* **2**, e132 (2006).
6. Fraumene, C., Petretto, E., Angius, A. & Pirastu, M. Striking differentiation of sub-populations within a genetically homogeneous isolate (Ogliastra) in Sardinia as revealed by mtDNA analysis. *Hum Genet* **114**, 1–10 (2003).
7. Francalacci, P. *et al.* Low-Pass DNA Sequencing of 1200 Sardinians Reconstructs European Y-Chromosome Phylogeny. *Science* **341**, 565–569 (2013).
8. Cucca, F. *et al.* The distribution of DR4 haplotypes in sardinia suggests a primary association of type I diabetes with DRB1 and DQB1 loci. *Human Immunology* **43**, 301–308 (1995).
9. Pugliatti, M. The epidemiology of multiple sclerosis in Europe. *European Journal of Neurology* **13**, 700–722 (2006).
10. Orrù, V. *et al.* Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell* **155**, 242–256 (2013).
11. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
12. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
13. Stranger, B. E. *et al.* Genome-Wide Associations of Gene Expression Variation in Humans. *PLoS Genetics* **1**, e78 (2005).
14. Göring, H. H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* **39**, 1208–1216 (2007).
15. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
16. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
17. Carlborg, Ö. *et al.* Methodological aspects of the genetic dissection of gene expression. *Bioinformatics* **21**, 2383–2393 (2004).
18. Schliekelman, P. Statistical Power of Expression Quantitative Trait Loci for Mapping of Complex Trait Loci in Natural Populations. *Genetics* **178**, 2201–2216 (2008).
19. Cheung, V. G. *et al.* Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Protocols* **33**, 422–425 (2003).
20. Kwan, T. *et al.* Heritability of alternative splicing in the human genome.
21. Veyrieras, J.-B. *et al.* High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genetics* **4**, e1000214 (2008).
22. Fraser, H. B. & Xie, X. Common polymorphic transcript variation in human disease. *Genome Research* **19**, 567–575 (2009).
23. Moffatt, M. F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
24. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**,

423–428 (2008).
25. Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nat Genet* **39**, 1494–1499 (2007).
26. Sieberts, S. Moving toward a system genetics view of disease - Springer. *Mamm Genome* **18**, 389–401 (2007).
27. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease : Article : Nature. *Nature* **452**, 429–435 (2008).
28. Dimas, A. S. *et al.* Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science* **325**, 1246–1250 (2009).
29. Holt, R. A. & Jones, S. J. M. The new paradigm of flow cell sequencing. *Genome Research* **18**, 839–846 (2008).
30. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat Meth* – (2013).doi:doi:10.1038/nmeth.2714
31. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Meth* (2013).doi:10.1038/nmeth.2722
32. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511–515 (2010).
33. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **5**, 621–628 (2008).
34. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31**, 46–53 (2012).
35. Simon Anders, W. H. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010).
36. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
37. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* **7**, 500–507 (2012).
38. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
39. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. **10**, 57–63 (2009).
40. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Meth* **5**, 613–619 (2008).
41. Muhammad A Tariq, H. J. K. O. J. N. P. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Research* **39**, e120 (2011).
42. Cui, P. *et al.* A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96**, 259–265 (2010).
43. Huang, R. *et al.* An RNA-Seq Strategy to Detect the Complete Coding and Non-Coding Transcriptome Including Full-Length Imprinted Macro ncRNAs. *PLoS ONE* **6**, e27288 (2011).
44. Kumar, V. *et al.* Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. *PLoS Genetics* **9**, e1003201 (2013).
45. Sanna, S. *et al.* Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. *Nature Protocols* **42**, 495–497 (2010).
46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
47. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2012).
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,

2078–2079 (2009).
49. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
50. Chen, W.-M. & Abecasis, G. Family-Based Association Tests for Genomewide Association Scans. *The American Journal of Human Genetics* **81**, 913–926 (2007).
51. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Protocols* **30**, 97–101 (2001).
52. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2011).
53. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178–192 (2013).
54. Nicol, J. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**, 2730 (2009).
55. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
56. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996–1006 (2002).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
58. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18**, 1435–1440 (2011).