UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Scienze Biologiche e Molecolari

XXVII Ciclo

BIOINFORMATIC TOOLS FOR NEXT GENERATION TRANSCRIPTOMICS

G.M. Prazzoli

PhD Thesis

Scientific tutor: Prof. Giulio Pavesi

Academic year: 2014-2015

SSD: BIO/11

Thesis performed at Università degli Studi di Milano.

INDEX

Abstract

In the last few years the introduction of novel technologies known as “next-generation sequencing” (NGS) has brought a major step forward in sequencing. These techniques have practically supplanted the conventional Sanger strategies that have been the principal method of sequencing DNA since the late 1970s. Different NGS platforms have been introduced, with the newest using ion-sensitive sensors to detect the incorporation of bases as opposed to the more commonly used fluorescent labelled nucleotides. Since the first techniques were introduced, both the sequencing runtime and the cost per sequenced base have dramatically decreased, and, at the current state of the art, a complete human genome can be fully sequenced in under 24 hours. On the other hand, the ever-increasing amount of short sequences (or reads) yielded per single run makes the processing of the data more difficult and challenging from a computational point of view. One of the most prominent and promising fields of application is RNA-Seq, an assay that provides a fast and reliable way to study transcriptomic variability on a whole-genome scale. Generally, in a RNA-Seq experiment, a RNA sample is converted in a cDNA library, which then undergoes several cycles of sequencing with a NGS method of choice. Usually, the resulting sequences are either mapped on the reference genome or assembled de novo without the aid of genomic sequence to produce a genome-scale transcription map, or trascriptome.

The data analyzed in this thesis comes from a three year research project focused on the characterization of tissue- and individual-specific alternative splicing, and its regulation. Data consist of several RNA-Seq experiments performed on different human tissues, coming from three healthy individuals. A total of 18 sets of data (6 tissues from three individuals with 3 replicates for each) were studied. The work initially focused on the quantification of mitochondrial DNA and RNA in the six individuals, and its variability. Then, we developed a computational method for the identification of tissue- and individual-specific transcripts, able to perform a multi-sample comparison. The algorithm we implemented employs statistical test based on a variant of Shannon’s information entropy, in order to identify transcripts with an expression pattern presenting a significant bias towards one or more of the samples studied. The results obtained show the method to be robust and efficient, overcoming the need of performing pairwise comparison as with the algorithms currently available, providing a thorough and complete map of the extent of tissue-specificity of gene expression at the single individual level.

State of the Art

RNA & gene expression

Euchariotic organisms store their hereditary information by encoding it in long deoxynucleotides molecules, which are packed and condensed in cell nucleus in the form of chromosomes (Watson & Crick, 1953). The single unit of DNA, or monomer, is known as nucleotide, a small molecule formed by a phosphatase group, a 5-carbon sugar (2'-deoxyribose) and a nitrogenous base called nucleobase. The latter has four different forms: adenine (A), timine (T, uracile for RNA), cytosine (C) and guanine (G). The nuclotides are bound to each other in two complementary strands (A with T and C with G) which are bound together forming hydrogen bonds, with a double helix shape. Thus both strands contain the same genetic information, but mirrored. Not all DNA contains encoded information, yet it is stored in shorter sequences scattered all over the DNA (genes), which represent the single unit of heredity.

The residues of the deoxyribose determine the DNA directionality: one end of a DNA polymer contains an expose hydroxyl group on the penthose, known as the 3' end while the other end contains a phosphate group, the 5' end. Directionality is essential to genes and gene expression, and the two strands are referred to with different names: the sense strand, which goes from 5' to 3' end, on which genes are traditionally annotated; the reverse strand which is the complementary opposite and goes from 3' to 5' end. This direction play an important role, since most of the processes involving DNA transcription or replication (Lehman, Bessman, Simms, & Kornberg, 1958) occur in a specific direction.

The central dogma of molecular biology (Figure 1), proposed by Francis Crick (Crick, 1958), states that the biological information encoded in DNA is transferred and translated into proteins, and this process is irreversible. More in depth, the information flow consists of three main stages: DNA, RNA and Protein. In the first stage, the information contained in segments of DNA (genes) is transcribed (i.e. copied) into a single strand ribonucleic acid molecule (second stage), known as RNA. This molecule, which is an exact copy of DNA can be either used as a template for translation into proteins or further processed as is, with other specific cell roles. According to this distinction, RNA can be divided into two main categories: coding RNA (mRNA, messenger RNA) or non-coding RNA (ncRNA).

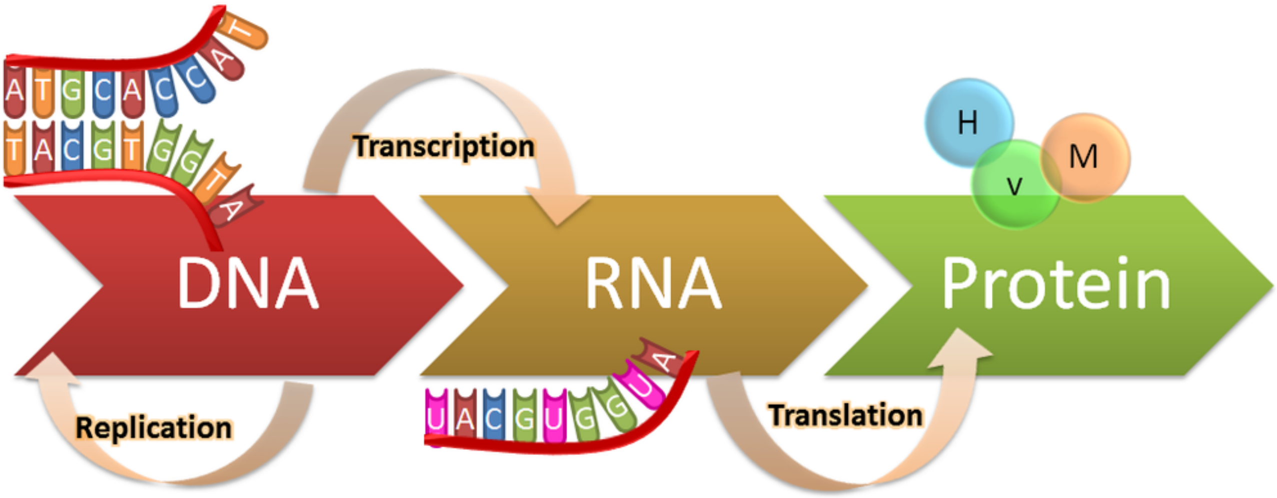


Figure 1. Visual representation of the central dogma of biology.

The process of copying DNA sequences into single-stranded RNA is known as transcription, and it is performed in the cell nucleus by an enzyme known as RNA polymerase, RNA pol (Furth, Hurwitz, & Anders, 1962). In euchariotes, different types of this class of enzymes exist, and they all synthetize different types of RNAs:

RNA pol I transcribes ribosomal RNA, which is combined with complexes known as ribosomes, structure on which the template mRNA is translated into amminoacid sequences (Russell & Zomerdijk, 2006);
RNA pol II transcribes mRNA, RNA bearing protein information (Kornberg, 1999)
RNA pol III transcribes tRNA, small RNA molecules whose role is to select specific amminoacids and incorporate them in ribosomes thus synthetizing proteins (Dieci, Fiorino, Castelnuovo, Teichmann, & Pagano, 2007).

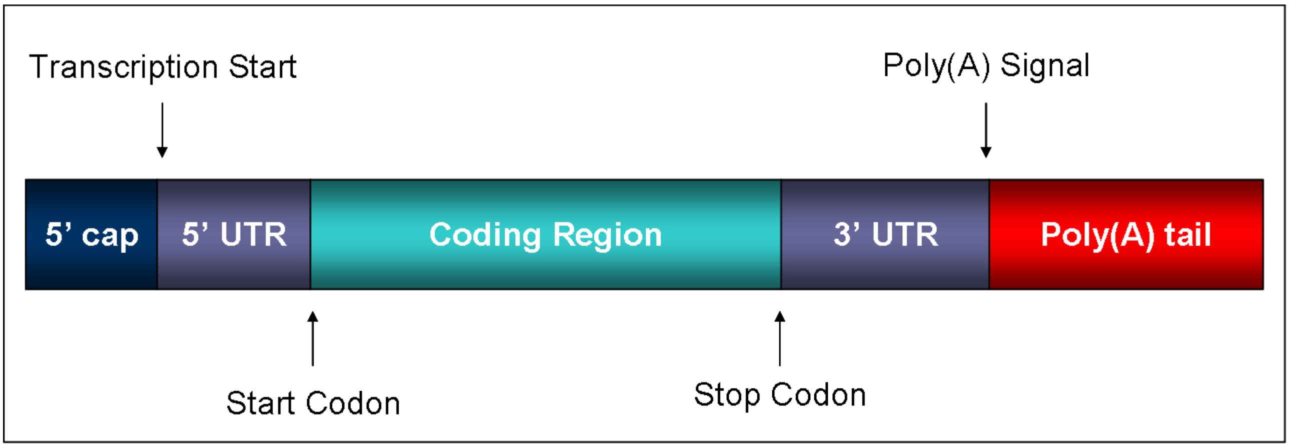


Figure 2. The typical structure of mRNA molecules.

The three polymerases further transcribe other small RNAs which play cathalitic and structural roles in the cell. The accepted model of RNA includes three main steps: initiation, elongation and termination. Initiation starts when RNA pol binds the DNA upstream of the gene to be transcribed at the promoter, with the aid of transcription factors (TF), proteins which assemble on the promoter and determine transcription levels. The DNA downstream the RNA pol starts to unbind the two strands for the RNa pol to read the gene sequence, and the complex of RNA pol advance adding nucleotides to the 3' end of the forming RNA complementary molecule (elongation). Once the gene has been completely copied and the stop codon is read the RNA pol leaves DNA, cleaving the RNA and releasing it from the transcriptional machinery (termination).

Post transcriptional modifications

Once released, the RNA undergoes several modifications, which are commonly referred to as post-transcriptional modifications: a series of processes that ensure RNA stability and facilitate its passage across the nucleus membrane. One of the most important modification is poly-A tailing (Colgan & Manley, 1997; Edmonds & Abrams, 1960): to stabilize RNA molecules, a series of adenine nucleuotides (hence the poly-A) is added to the 3' end of RNA by the poly-A polymerase. Generally, ~200 nucleotides are added, although there are some exceptions, for example the iston protein-coding mRNA gets no poly-adenilation (Davila Lopez & Samuelsson, 2008). The tail length ensures that the RNA is degradated accordingly, therefore a longer poly-A tail means that the transcript will be translated more proactively.

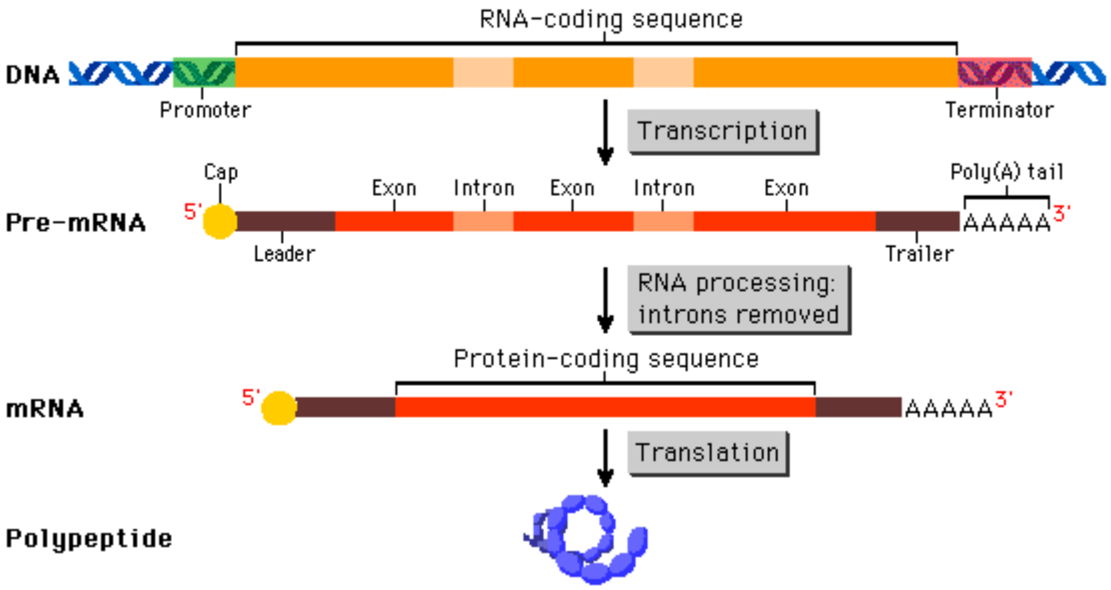


Figure 3. The flow from gene to pre-mRNA to mature RNA

Another important process is splicing (Black, 2003): the RNA in the nucleus is in fact known as pre-RNA for all the modifications it undergoes. The RNA is basically formed by alternated exons and introns, the first being (with some exceptions) the coding part of the RNA, the latter being the non-coding part. When the RNA undergoes splicing, the introns are excided and flanking exons are bound together, and the resulting molecule is called mature RNA. Chemically speaking, the introns excisions consist in two sequential transesterification reactions, which break and join back together the phosphodiesterical bonds between nucleotides on pre-mRNA, by acomplex protein machinery called spliceosome. Spliceosome is about the same size as a ribosome and is formed by ~150 proteins and 5 RNA molecules (small nuclear RNAs, U1, U2, U4, U5, U6), which recognize splicing sites on RNA and catahlize the intron excision. The snRNAs couple with proteins of the spliceosome forming snRNP (small nuclear ribonuclear proteins) which play different key roles during the whole splicing process, which can be summarized in three main phases (Figure 4):

1) In the first phase, the U1 snRNP identifies and hybridizes the 5' intron splicing site (known as donor site) on the pre-mRNA. The donor site includes an invariant GU sequence in a less conserved and broader region. Conversely the splicing site at 3' end (known as branching site) contains a highly conserved AG sequence, recognized by snRNP U2. The GU and AG seuqncences define the start and end of the intron respectively. U4, U5 and U6 join the oher two RNA-protein complexes, and loop the pre-mRNA, getting the GU and AP sites near each other.

2) In the second phase, the spliceosome complex undergoes a conformational modification, creating a second ary RNA structure called lariat, circularizing the intron.
3) The lariat is cleaved and the flanking exons are bound together.

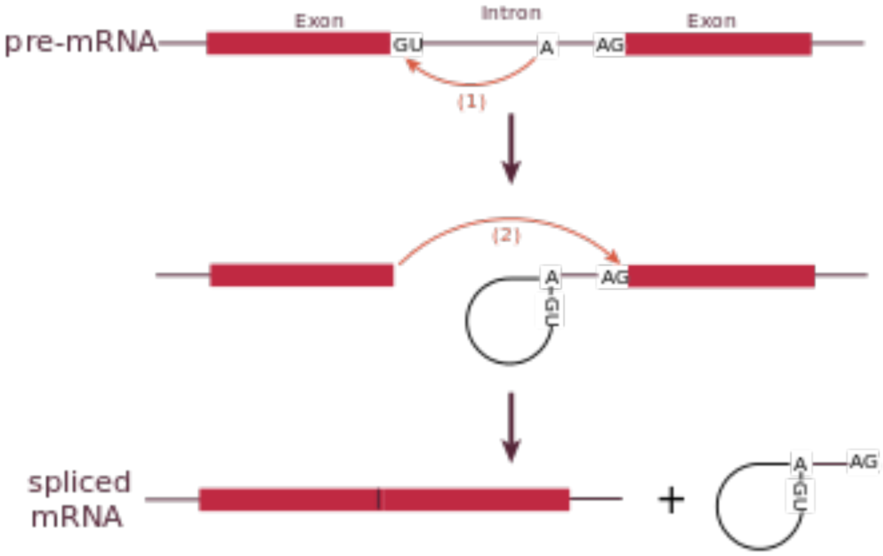


Figure 4. Visual representation of the splicing process.

An interesting feature of splicing is alternative splicing: since the discovery of DNA, it was common knowledge that each gene would give origin to a single RNA and a single protein. Gene regulation was in fact first studied in relatively simple bacterial systems. Most bacterial RNA transcripts do not undergo splicing, being colinear, with DNA directly encoding them. That assumption, anyway, could not explain why complex organism like humans had a smaller genome size than some plants or other simpler organisms: a wider genome size would mean more genes and therefore more proteins for more complexity. However this 1:1 ratio dogma was proven wrong when alternative splicing was described for the first time: In 1977, several groups of researchers who were working with adenoviruses that infect and replicate in mammalian cells obtained some surprising results. These scientists identified a series of RNA molecules that they called "mosaics," each of which contained sequences from noncontiguous sites in the viral genome (Berget & Sharp, 1977; Chow, Gelinas, Broker, & Roberts, 2000). Those sequences became lately known as exons and introns.

Alternative splicing is the process by which exons and introns are rearranged in different ways, by different excision mechanisms, originating different transcripts from the same primary transcripts. Recent transcriptome studies (Pan, Shai, Lee, Frey, & Blencowe, 2008) report that more than 95% of the total multiexon human genes can undergo alternative splicing. The usage of a particular splice site is a key factor in determining the relative abundance of a specific isoforms. For this reason, the entire process of splicing (both alternative and non-alternative) is regulated by trans-acting proteins (called activators/repressors) that bind to cis-regulation sites located on the pre-mRNA, selectively enhancing or silencing the usage of splice sites (Barbosa-Morais et al., 2012).

Alternative splicing comes in many different forms (Matlin, Clark, & Smith, 2005; Pan, Shai, Lee, Frey, & Blencowe, 2008; Sammeth, Foissac, & Guigo, 2008), as shown in Figure 5:

Exon Skipping: as the name suggests, an exon is excluded from the mature RNA by excision of the flanking introns. This is the most common and rearranging effective type of splicing in mammals.

Mutually exclusive exons: This happens when the isoforms can present just one of a set of two exons in the different isoforms produced, but not both at the same time, as the name implies.

Alternative 5’ (or 3’) end sites: an alternative splice site is used, altering the 5’ (3’) end of an exon.

Intron retention: while the mature RNA is said to contain no introns, retained intron alternative splicing actually keeps an intron in the mature RNA. This can happen if splicing sites are not effectively recognized by the spliceosome machinery. Retaining of an intron can cause frameshift in the codons, breaking the reading frame and resulting in defective or incorrectly folded proteins.

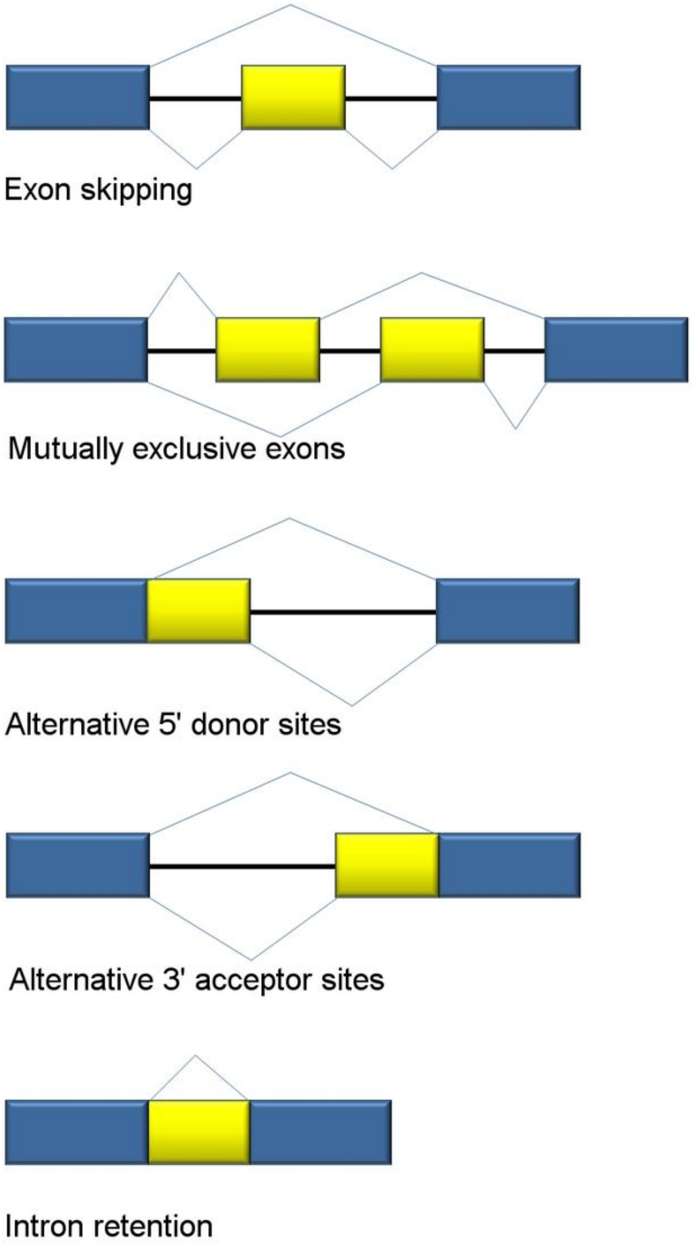


Figure 5. The five main types of alternative splicing events.

DNA Sequencing

Since its discovery in 1953 (Watson & Crick, 1953), one of the biggest challenge to scientists has been "cracking the code" of DNA, that is, deciphering DNA sequences and assign them their functional role in the cell. Since the advent of the first capillary electrophoresis method, CE-Sanger sequencing (Sanger, Nicklen, & Coulson, 1992), scientists from all over the world have tried to elucidate and shed light on genetic information and its mechanism. Although this method has been widely used and has been the de-facto standard for DNA sequencing for many decades, it is heavily hampered by many limitations, such as sequencing speed, scalability (sequencing whole genomes took years of work), and strict laboratory protocols.

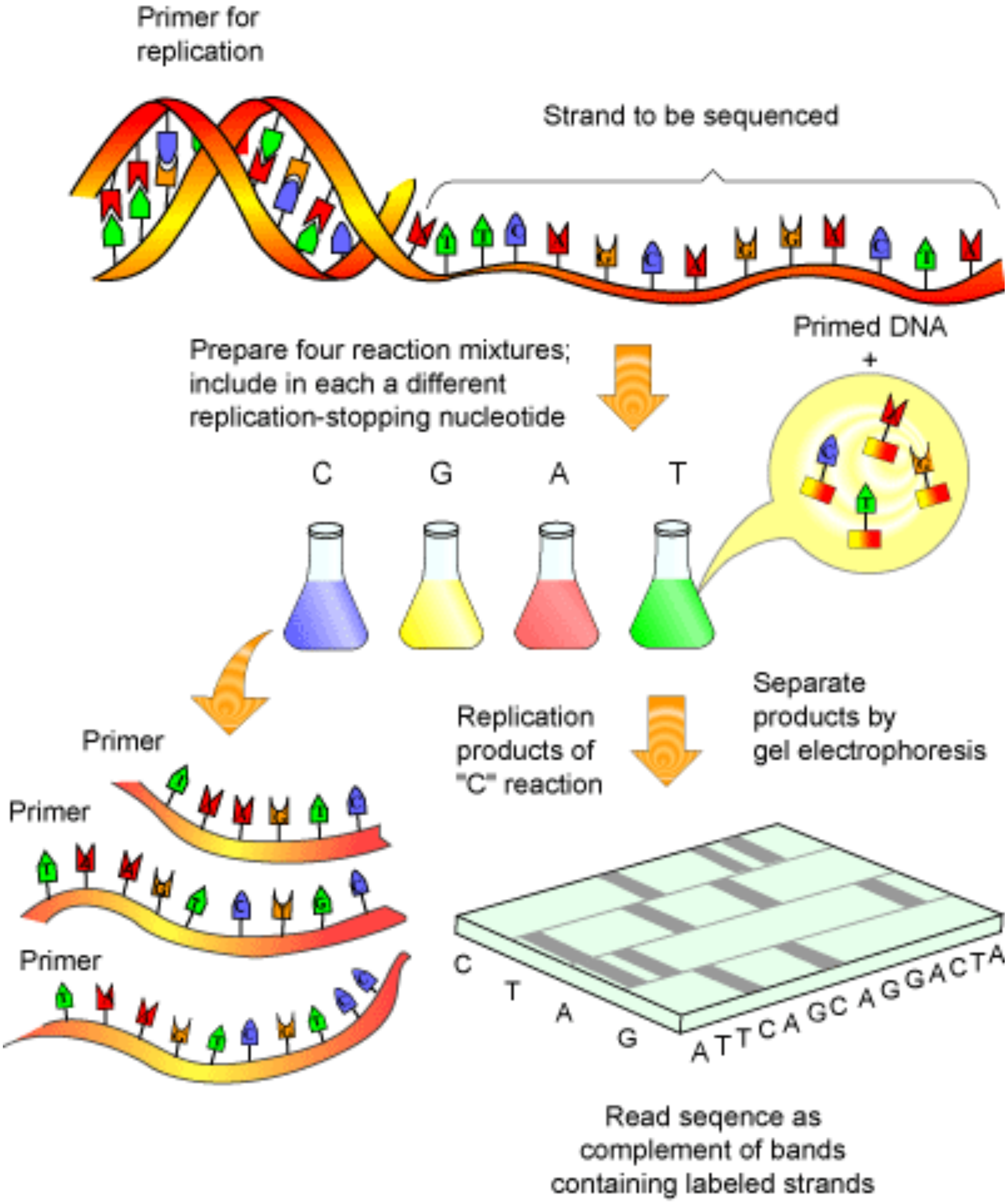

Primer for replication
Strand to be sequenced
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA +
C
G
A
T
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
C
T
A
G
ATTCAGCAGGACTA
Read seqence as complement of bands containing labeled strands

Figure 6. In th eclassical formulation of the Sanger sequencing method, 4 radioactive terminator nucleotides are used. When incorporated by DNA Pol, these nucleotides stop DNA synthesis and produce truncated DNA fragments of different lengths, that can be separated with the aid of gel electrophoresis allowing for sequence reconstruction.

To overcome these limitations, a groundbreaking new technologies were developed and introduced in the early years of 2000, later on called Next Generation Sequencing (NGS, (Marziali & Akeson, 2001)). NGS introduced a completely new way of sequencing, which could process million of bases in the span of few hours keeping anyway a high precision level. In principle, the concept behind several different NGS methods can be considered similar to the Sanger method: the sequence is determined by the incorporation of specifically modified nucleotides in a fragment of DNA re-syntetized from a template strand. Though, in contrast to Sanger method, the type of signal is different: almost all NGS techniques rely on light emitted by incorporated nucleotides, each labelled with a specifc fluorescent marker. This technology allows the processing of many different DNA fragments in a parallel fashion, producing hundreds of Gigabases (Gb) in a single run.

Since its first appearance in 2004, many new platforms based on different approaches have been developed, yielding an ever-increasing data output: roughly, each year the data produced by a single machine run has practically doubled and the trend has not stopped ever since, with new, more performant technologies developed each year (Niedringhaus, Milanova, Kerby, Snyder, & Barron, 2011). For example, while in 2007 a single sequencing run averaged 1 Gb of data, in 2011 the rate reached near 1Tb, an astonishing 1000x increase in output amount. Even the per-sequenced-base cost has been decreasing since the first technologies came out, and today a whole human genome can be processed in a few hours at an average cost of 30,000-50,000$; for comparison, the human genome project started in 1990 (Sawicki, Samara, Hurwitz, & Passaro Jr, 1993; Venter et al., 2001) took more than 10 years (including bioinformatic analysis and assembly) and required almost 3 billion dollars founds to complete. The results were published in 2003, just one year before NGS was developed.

Even though NGS methods are high-throughput, they are easily scalable to the needs of scientists: not only a whole genome can be sequenced, but smaller samples of DNA, or RNA selected according to some criterion.. Moreover, the sequencing depth can be fine-tuned, tailoring it to the needs of the experiment, adjusting the coverage and the average length of the reads. Furthermore, different experiments can be sequenced in the same run, a technique called multiplexing, which uses specific DNA adapters called barcodes, in order to differentiate the sequences in the data analysis (Illumina, 2015).

454 GenomeSequencer

First technique to be commercially available (2004-2005), produced by Life Sciences (now Roche) (Roche, 2015), is called 454. The method is based on pyrosequencing, that is, pyrophosphate detection (Margulies et al., 2005; Ronaghi, Karamohamed, Pettersson, Uhlen, & Nyren, 1996; Ronaghi, Uhlen, & Nyren, 1998). Library construction is performed by nebulization, in order to shear DNA into small fragments, and a subsequent enzymatic blunt-ending and/or adapter ligatition. In pyrosequencing, every nucleotide incorporation, performed by DNA polymerase, is coupled with pyrofosfate emission, which triggers a signal process. This results in light (in the range of visible spectrum) being emitted by luciferase enzyme: luciferase is the generic name given to a class of enzymes that play key-roles in bioluminescence processes. The most important luciferase enzyme is firefly luciferase, found in the species Photinus pyralis (Nyren, 2007). The quantity of light emitted is proportional to the number of incorporated nucleotides. For this approach, the DNA fragments library (ligated with a specific linker) is put into a solution containing many agarose beads: the 454 specific adapters contain a 5’ biotin tag, which allows binding on streptavidin coated beads. A high enough dilution ensures that most of the beads are bound to one single fragment of the library to be sequenced. Each of these complexes is then isolated in a water micelle in an oil solution which contain the necessary reagents to start the PCR amplification (Mullis et al., 1986). The beads are then put in a grid with 44um wells, sufficient to to contain a single bead. Over the grid, an high-resolution optical sensor register every light emission, 400.000 wells in parallel. The first of the 4 nucleotides (TCGA) is injected in the support, triggering, where incorporated, the luciferase reaction. This strategy allows to calibrate the machine to a single nucleotide precision, and to have good accuracy up to 6 single nucleotides incorporated in series (Gilles et al., 2011). Furthermore, because every step uses a single base type, base calling errors are practically absent. Performance-wise, this method can sequence up to 100 Mb in 8hours, which correspond roughly to 100 nucleotide injections. Since its debut, the read length has increased from 250 nt to 500+ nt, and it is still the method of choice for de-novo genome assembly.

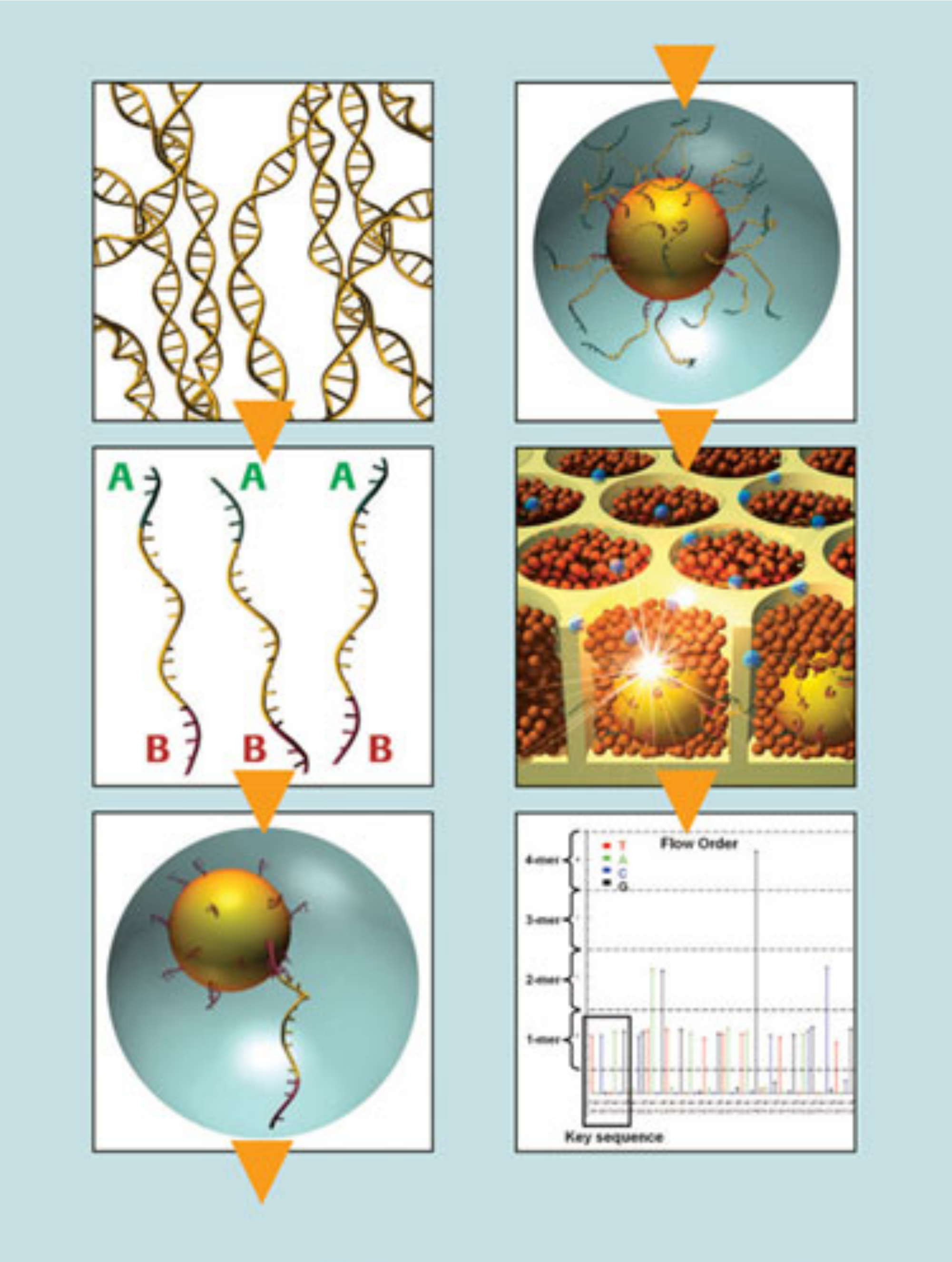


Figure 7 The workflow for 454 sequencing

Ilumina Solexa

The “sequencing-by-synthesis” technology now used by Illumina (Illumina, 2015) was originally developed by Shankar Balasubramanian and David Klenerman at the University of Cambridge (Bentley et al., 2008). They founded the company Solexa in 1998 to commercialize their sequencing method. Illumina went on to purchase Solexa in 2007 and has built upon, and rapidly improved the original technology.

The Solexa/Illumina sequencing method is similar to Sanger sequencing, but it uses modified dNTPs containing a terminator which blocks further polymerization (Bentley et al., 2008; Turcatti, Romieu, Fedurco, & Tairi, 2008).

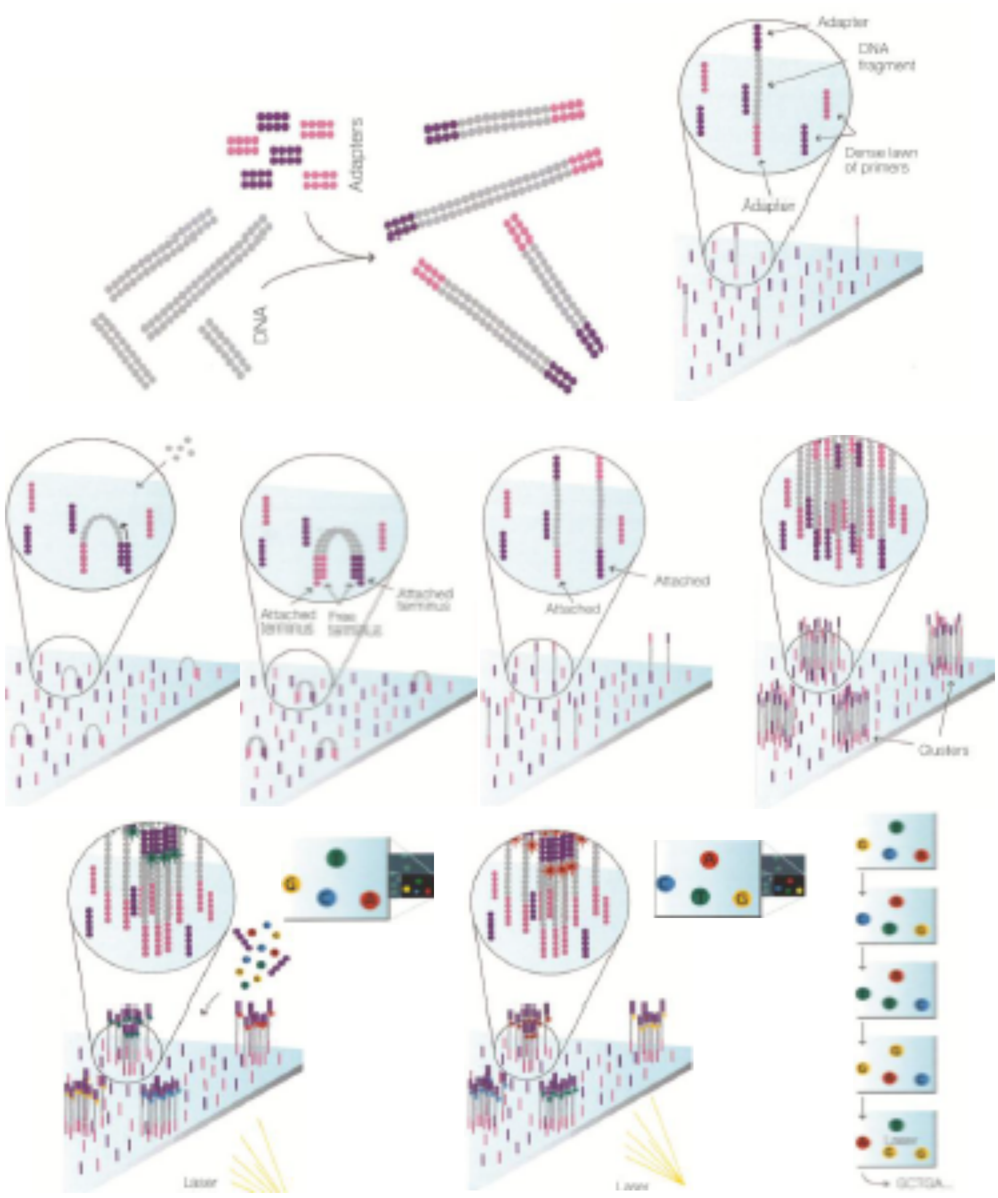


Figure 8: The workflow of Illumina sequencing

Although the fluorescent imaging system used in Illumina sequencers is not sensitive enough to detect the signal from a single template molecule, the major innovation of the Illumina method is the amplification of template molecules on a solid surface (Adessi et al., 2000; Fedurco, Romieu, Williams, Lawrence, & Turcatti, 2006). The DNA sample is prepared into a “sequencing library” by the fragmentation into pieces each around 200 bases long. Custom adapters are added to each end and the library is flowed across a solid surface (the “flow cell”) and the template fragments bind to this surface. Following this, a solid phase “bridge amplification” PCR process (cluster generation)

creates approximately one million copies of each template in tight physical clusters on the flowcell surface. Illumina has improved its image analysis technology dramatically which allows for higher cluster density on the surface of the flowcell.

Solid phase amplification is employed to create up to 1,000 identical copies of each single molecule in close proximity (diameter of one micron or less) (Adessi et al., 2000).

Solexa sequencing uses four proprietary fluorescently-labeled modified nucleotides to sequence the millions of clusters present on the flow cell surface. These nucleotides, specially designed to possess a reversible termination property, allow each cycle of the sequencing reaction to occur simultaneously in the presence of all four nucleotides (A, C, T, G). In each cycle, the polymerase is able to select the correct base to incorporate, with the natural competition between all four alternatives leading to higher accuracy than methods where only one nucleotide is present in the reaction mix at a time, like 454. Sequences where a particular base is repeated one after another (e.g., homopolymers) are addressed like any other sequence and with high accuracy.

SOLiD Sequencer

The technology behind the SOLiD™ (Biosystems, 2015) (Sequencing by Oligo Ligation and Detection) platform was first described in 2005 by Shendure (Shendure et al., 2005). The first machine was commercially sold in October 2007 (Mardis, 2008). An overview of the workflow of SOLiD sequencing is shown in Figure 1-5. Library construction is similar to both 454 and Illumina, and may be constructed with different approaches to produce adaptor-flanked fragments (Shendure & Ji, 2008). The SOLiD technology also requires DNA amplification, and like 454 sequencing this is performed using emulsion PCR (Dressman, Yan, Traverso, Kinzler, & Vogelstein, 2003) where DNA fragments are bound to paramagnetic beads. Prior to sequencing, the emulsion is broken, and beads enriched and immobilized to the surface of a specially treated glass slide (Mardis, 2008), generating a dense array. After hybridization of a sequencing primer, the synthesis of DNA is not performed by a DNA-polymerase, instead by a ligase (Housby & Southern, 1998; Shendure et al., 2005). At each step, a fluorescently labeled octamer originating from a degenerate set is ligated to the DNA fragment. Fluorescent markers bound to the octamers are correlated to a specific position within the oligo. After image acquisition in four different channels, chemical cleavage of the octamer between the fifth and sixth base is performed removing the marker. Multiple ligation steps enable sequencing of every fifth base of the DNA fragment. Following several rounds of ligation, image acquisition and cleavage, the DNA is denatured, enabling annealing of a new sequencing primer at a different position on the adaptor sequence. A unique feature is that the fluorescent markers are correlated to dinucleotides, and not just a single base. This combined with an alternate use of sequencing primers and octamer sets, where the fluorophores correspond to different positions on the octamer, ensures that each base is sequenced twice, and thus minimizing base calling errors (Heinz, 2010; McKernan et al., 2008).

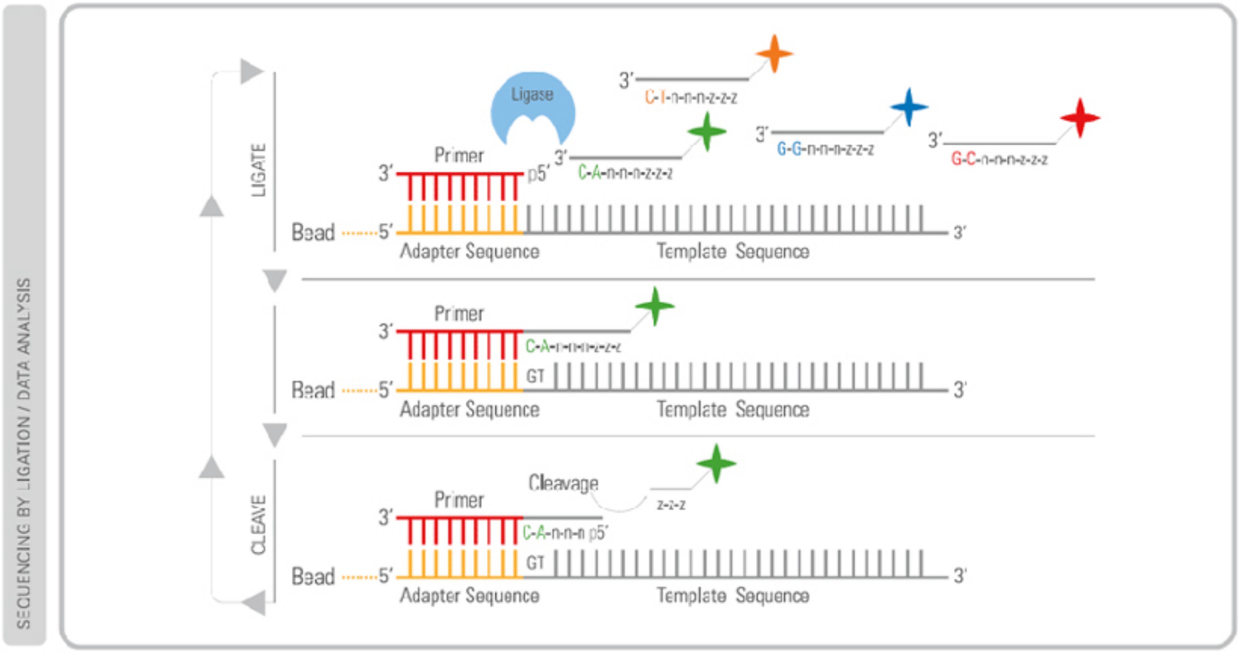


Figure 9. The workflow of Solid sequencing

Paired-end Sequencing

The three platforms described above are more or less (let alone 454) limited by short read lengths. However, this limitation has been partly overcome by the development of paired-end sequencing (Roach, Boysen, Wang, & Hood, 1995), which can be performed using all three sequencing systems. Moreover, this kind of sequencing can be used for more in depth analysis, allowing insertion/deletion detection and exon excision to start with. Paired-end reads (sometimes referred to as mate reads) are short sequences originating from both the two 5' ends of a DNA fragment (Fullwood, Wei, Liu, & Ruan, 2009). Paired-end sequencing was already described in 1981 by Hong (Hong, 1981), and the first use of paired-end sequencing was reported by Ewards and Caskey in 1990 (Edwards & Caskey, 1991). There are multiple ways of constructing a paired-end library:

1) Clone based method, in which the target DNA sequence is linked with adaptors containing MmeI restriction sites immediately next to the sequence which needs to be sequenced. Following amplification in E. coli, purification and MmeI digestion, the tag containing vector is recircularized, which results in joining the two sequence tags. After amplification in E. coli, the circularized constructs can be purified using restriction digestion (Ng, Wei, & Ruan, 2007).
2) A second method was introduced by Shendure (Shendure et al., 2005): the target DNA fragments are directly circularized with linker oligonucleotides hereby joining the two ends of the target DNA. The linker sequence contains two restriction sites (e.g. MmeI) flanking the two ends of the target DNA, enabling restriction digestion to release the tag-linker-tag construct for sequencing (Shendure et al., 2005). These two methods can create libraries with long inserts (up to 20 kb) between the two sequence tags, which are often referred to as mate pair libraries (Fullwood et al., 2009). Additional to these methods, short insert libraries (200-500 bp) can also be paired-end sequenced using Illumina sequencing. Here paired-end libraries are made using adaptors with two different sequencing primers. Paired-end is performed by first sequencing the target DNA utilizing the first sequencing primer. After subsequent product denaturation, bridging, and second strand synthesis, the opposite strand is cleaved providing a template for a round second sequencing utilizing the second sequencing primer (Bentley et al., 2008) .

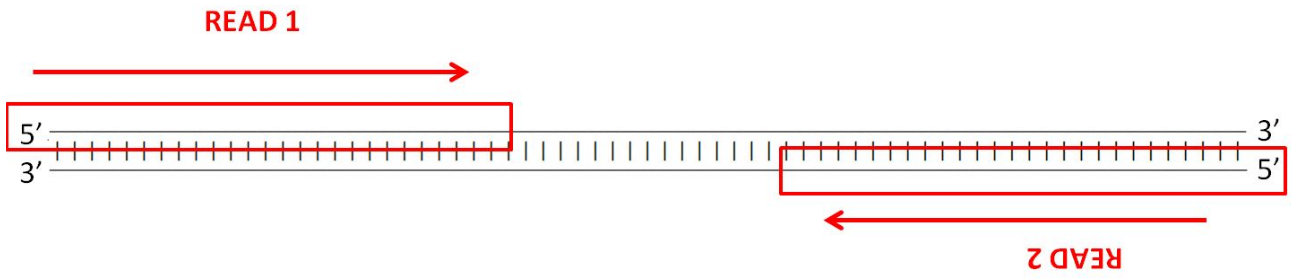


Figure 10. Paired end reads

Directional sequencing

Another feature of NGS is the possibility to keep track of reads directionality, that is, retain information about which strand the reads come from, by using kits that allow directional library construction. In a typical NGS-based experiment, genes/transcripts levels are computed by total read enrichment, without preserving strand information; consequently, sequencing data show equal levels of enrichment from both strands.

As an example, consider a generic gene A to be located on forward (+) strand of a chromosome. In a non-directional RNA-Seq experiment, gene A's exons are enriched by an equally distributed mixture of '+' and '-' strand-originated reads; conversely, by using directional sequencing all reads would come from the forward strand.

When trying to investigate expression levels of overlapping/proximate genes located on different strands or estimating antisense transcription rates, the aid of directional sequencing can help remove read mapping ambiguity and allows correct reads assignment (Flaherty, Van Nieuwerburgh, Head, & Golden, 2011).

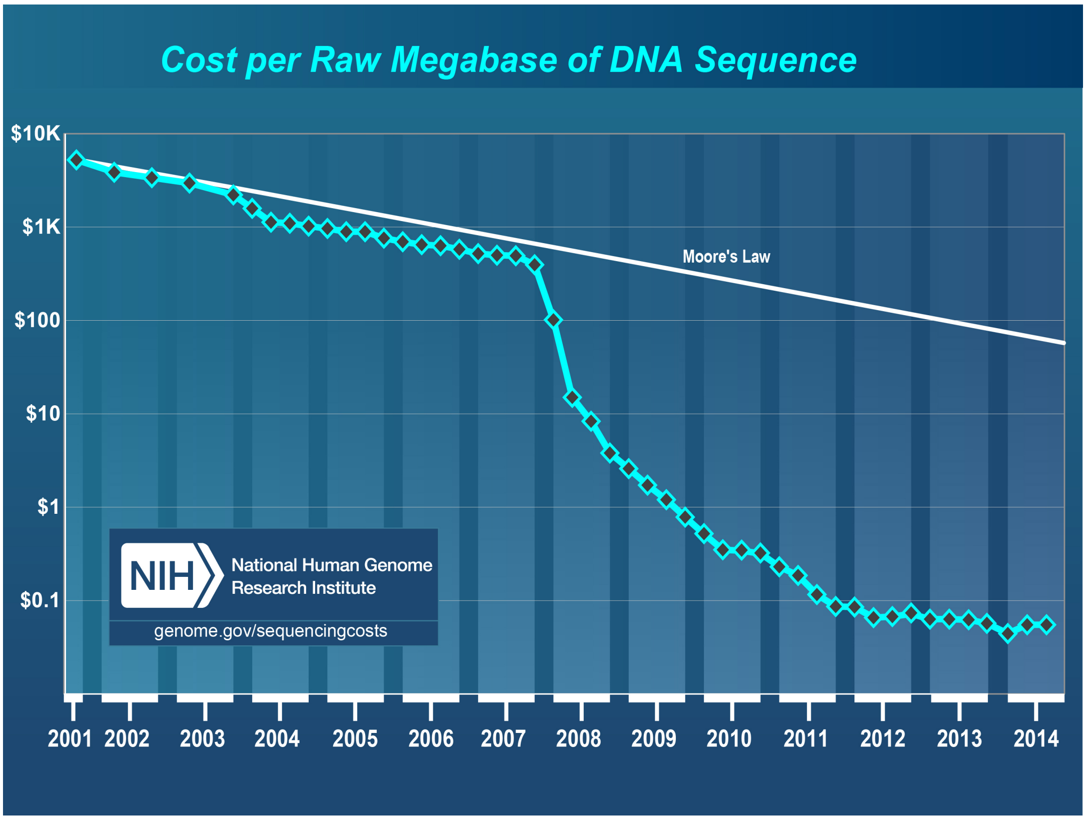
When dealing with paired end reads, one must bear in mind that only read pairs with proper orientation (i.e +/-; +/+ and -/- pairs can occur due to sequencing errors or sequence artifacts created during ligation/amplification processes) should be taken into account to correctly estimate enrichment.

Next generation sequencing costs

The National Human Genome Research Institute (NHGRI, 2015) compile on a yearly basis a “cost per genome” table to illustrate the overall cost of sequencing an entire human genome, by gathering data from its sequencing centers (Wetterstrand, 2015; Figure 11a and 11b). Calculations take into account labor, amortization of sequencing machine cost, computational postprocessing and sample preparation. Table 1 illustrates the decreasing cost trend to sequence a complete human genome (approximately 3 Gb) over time, since the first sequenced human genome was released in 2001. The dramatic drop in costs dating back to 2008 is the result of the transition from classic sequencing methods (Maxam & Gilbert, 1992; Sanger et al., 1992) to NGS technologies. NGS methods produce shorter sequences and require a greater sequencing depth to allow correct genome assembly, however, high throughput parallel sequencing reduces resources cost and the number of sequencing runs, hence decreasing overall costs.

	Read Length	**Sequence per day**	**Cost per base**
454	400-500	1 Gb	18 $/Mb
Illumina	100-150	6.5 Gb	0.4 $/Mb
Solid	50-80	5 Gb	0.5 $/Mb

Table 1. Read length, sequence per day and cost per base statistics of the three main NGS platforms



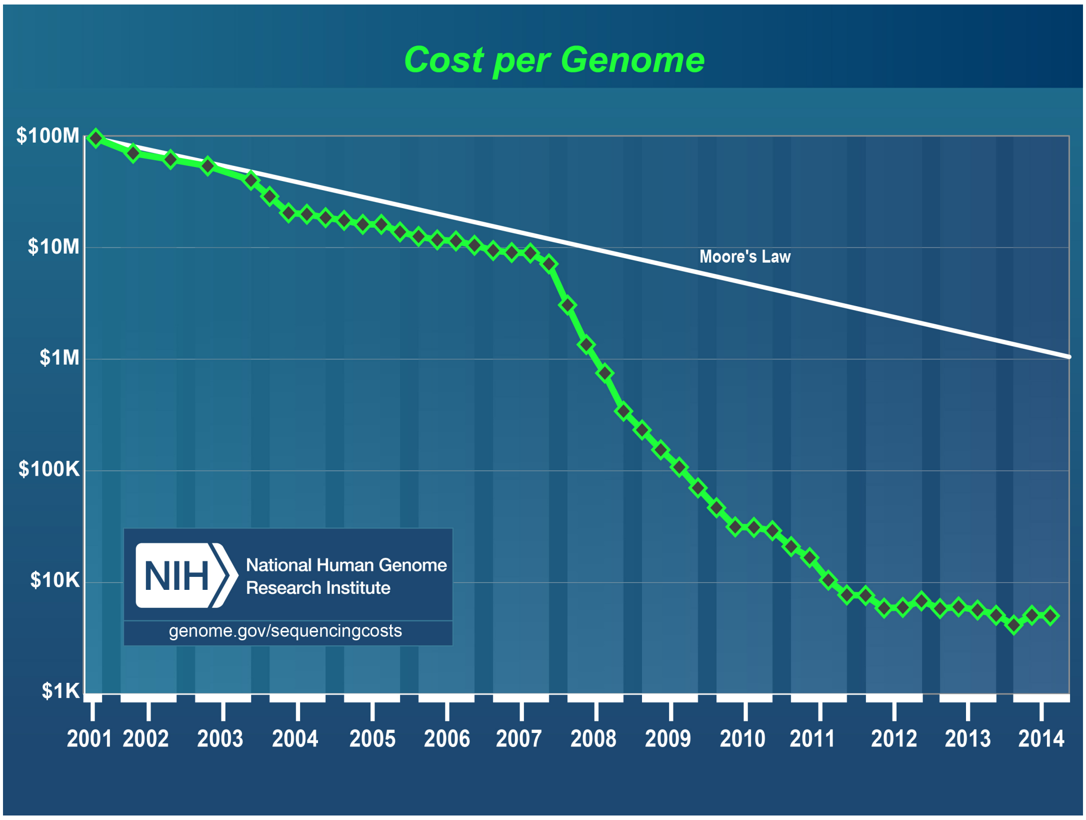


Figure 11a and 11b. The National Human Genome Researcg institute compiled two graphs by gathering data on sequencing costs in the last two decades. The results show that overall costs (calculated on the sequencing of 3000 Mb human genome) not only decrease in time, but overstrip Moore's law curve by a great margin.

The future of sequencing

Although PCR amplification (Adessi et al., 2000; Mullis et al., 1986) has revolutionized DNA analysis, in some instances it may introduce biases such as base sequence errors or favor the amplification of certain sequences over others, thus modifying the relative abundance of DNA fragments that existed before amplification. More accurate results could be achieved if a nano-scale sequencing of the starting DNA population without the aid of amplification steps would be possible (Schadt, Turner, & Kasarskis, 2010). Hence, third generation sequencing methods have been focusing on single-molecule sequencing, trying to achieve better accuracy.

Transcriptomics

In cellular and molecular biology everything ending with –ome have the sense “all of the specified constituents of a cell, considered collectively or in total”, and thus the -omics suffix indicates the measurement of an entire collection of biological molecules or sequence information. It is fair to say that all of the “-omics” approaches in the biological field, born to quantify and classify biological data, stem from the advance of technology (Schneider & Orchard, 2011). In the field of biology, 4 different “omics” branches have emerged in the past years:

1) Genomics, the quantitative and qualitative study of one cell DNA and genes, with the aim of assigning to each gene its functional role in the cell.
2) Transcriptomics, the quantitative and qualitative study of one cell total RNA, extending from messenger RNA (mRNA) to non-coding RNAs (ncRNA). The cataloguing of new RNA is still ongoing and it is still one the hottest topics of next generation “omics”
3) Proteomics, the quantitative and qualitative study of the whole collection of proteins encoded by genes (Kalia & Gupta, 2005);
4) Metabolomics, the quantitative and qualitative study of metabolites

These four main areas branched in many sub-topics, such as pharmacogenomics, i.e. the investigation of how and why differences in genomes lead to different drug responses and individual-tailored drug development (Daly, 2010), nutrigenomics, which focus on how different foods can trigger the insurgency of diseases by interacting with genes (Mead, 2007), to cite some. Since most, if not all of these disciplines require huge amount of data to be generated and processed, it is only natural that NGS techniques were the methods of choice for the biology "omics" era. As such, the development of NGS has brought the study of genome and transcriptome to a whole other level, and has lead to many important milestone discoveries, enhancing our understanding of the mechanism behind several human diseases and their pathogenesis (Rogers & Venter, 2005; Zhang, Chiodini, Badr, & Zhang, 2011). However, this paradigm shift in data collection and analysis has introduced new challenges, and the stunning benefit of NGS techniques are almost impossible to achieve without proper bionformatic workflows and frameworks (Stein, 2011).

Transcriptomics in particular represents the inevitable result of endless efforts to quantify the gene expression levels of thousands of genes in a parallel fashion since the discovery of DNA and RNA. It was not until late 90s (with the first appearance of microarrays) that scientists, having to face an enormous quantity of data, decided to abandon low-troughput approaches to embrace new methods. By the term transcriptome, it is intended the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition (Adams, 2008). Understanding the transcriptome mechanism is essential for analyzing the functional elements of the genome, cell development and genetic diseases. Similar to genomics for DNA and genes, the aim of transcriptomics is to obtain a complete expression profile of a cell, sequencing and cataloguing all the transcripts, assigning them their functional role in the cell, quantifying their relative abundance in certain conditions, determine the transcriptional structure of the underlying genes, defining their starting sites, 5' and 3' ends, splicing patterns and last but not least post transcriptional modifications. This of course includes all possible types of RNA: albeit in the beginning of the genomic era it was common knowledge that most of the RNA in the cells were composed by mRNA (messenger RNA, which gets polyadenilated, spliced and translated into protein), many new types of this one-stranded molecule were found in the last years, ranging from non-coding RNAs (ncRNAs) to silencing RNAs (siRNAs). The ever increasing discovery of new RNA types switched the focus from genomes to transcriptomes, and many new techniques were developed to study these molecules. In fact, one of the biggest challenges of RNA compared to DNA is the high degradation rate of RNA, which needs specific and sophisticated lab techniques to be isolated, purified and sequenced.

Non NGS-Based Technologies

Various methods and technologies have been implemented through the years to analyze RNAs and quantifying transcript abundance. Hybridization approaches (Lipshutz, Fodor, Gingeras, & Lockhart, 1999; Shalon, Smith, & Brown, 1996) make use of fluorescent markers applied to cDNA sequences (i.e. retro-transcribed RNA fragments, converted to double-strand DNA to avoid degradation and secondary structure formation, such as hairpin loops) which are then hybridized to high-density oligo microarrays, designed in order to provide a unique probe specific for each transcript of the genome studied. The intensity of the fluorescence on a specific spot then determines the relative abundance of a specific transcript. Although many types of microarrays have been developed in the meantime, like specialized microarrays with probes spanning exon-exon junctions sequences in order to detect splicing isoforms (Johnson et al., 2003), or tiling microarrays, a high density array which includes the whole genome sequence, which allows for a better resolution, this method reliance on a pre-existing genome/transcriptome sequence and a limited dynamic range of detection which suffers from background noise and signal saturation made scientists search for other methods (Shendure, 2008). Moreover, microarrays made difficult to compare transcript levels across multiple experiments, due to analogic signal quantification biases (Draghici, Khatri, Eklund, & Szallasi, 2006).

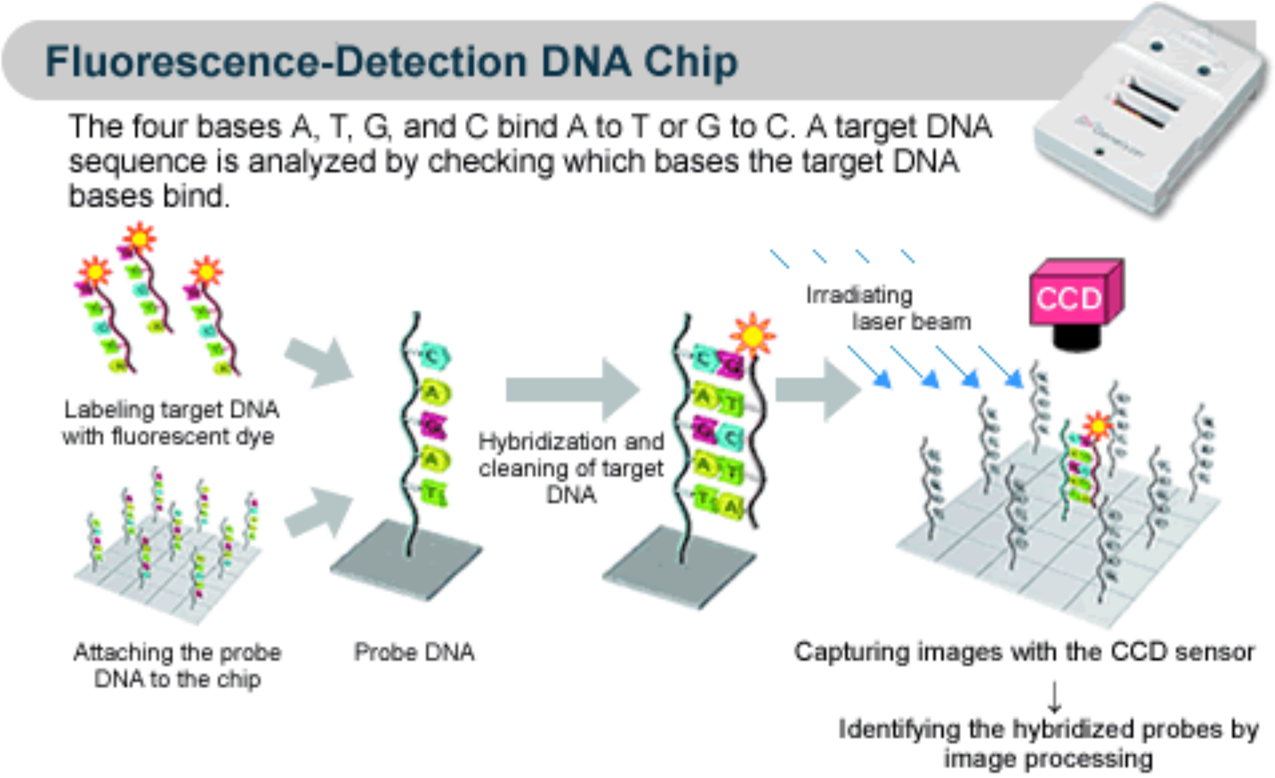

Fluorescence-Detection DNA Chip
The four bases A, T, G, and C bind A to T or G to C. A target DNA sequence is analyzed by checking which bases the target DNA bases bind.
Labeling target DNA with fluorescent dye
Attaching the probe DNA to the chip
Probe DNA
Hybridization and cleaning of target DNA
Irradiating laser beam
CCD
Capturing images with the CCD sensor
Identifying the hybridized probes by image processing

Figure 12. The basic wokflow for microarray assays.

Conversely, methods based on sequencing do not rely on existing sequences but directly determine them instead. Initially, Sanger sequencing of cDNA or EST libraries was used, but this approach suffers from low throughput and is not generally quantitative, therefore transcripts levels cannot be accurately measured. To overcome this restriction, Tag-based approaches were developed, such as serial analysis of gene expression (SAGE) (Velculescu, Zhang, Vogelstein, & Kinzler, 1995), cap analysis of gene expression (CAGE) (Shiraki et al., 2003) and massively parallel signature sequencing (MPSS) (Brenner et al., 2000). However, these are expensive methods and most of the tags produced map ambiguously on the reference genome. The SAGE technology produces sequence tags from mRNA transcripts, and, conversely to microarray, it gives a digital quantification of the mRNA transcript abundance (Velculescu et al., 1995). After cDNA conversion and binding to streptavidin beads, sequences are digested with the restriction enzyme NlaIII, which recognizes the site 5'-CATG-3'. Following ligation of a linker that contains a recognition site of the Type IIS restriction endonuclease BsmFI, the fragment is cleaved 15 bp in the 3′ direction from the recognition site, releasing the sequence tag. After removal of the linker fragment, tags are concatenated, cloned into a plasmid vector, and sequenced using Sanger method. SAGE and newer SAGE assays like superSAGE libraries usually contain between 10 and 100 thousand tags (Matsumura et al., 2005; Matsumura et al., 2010).

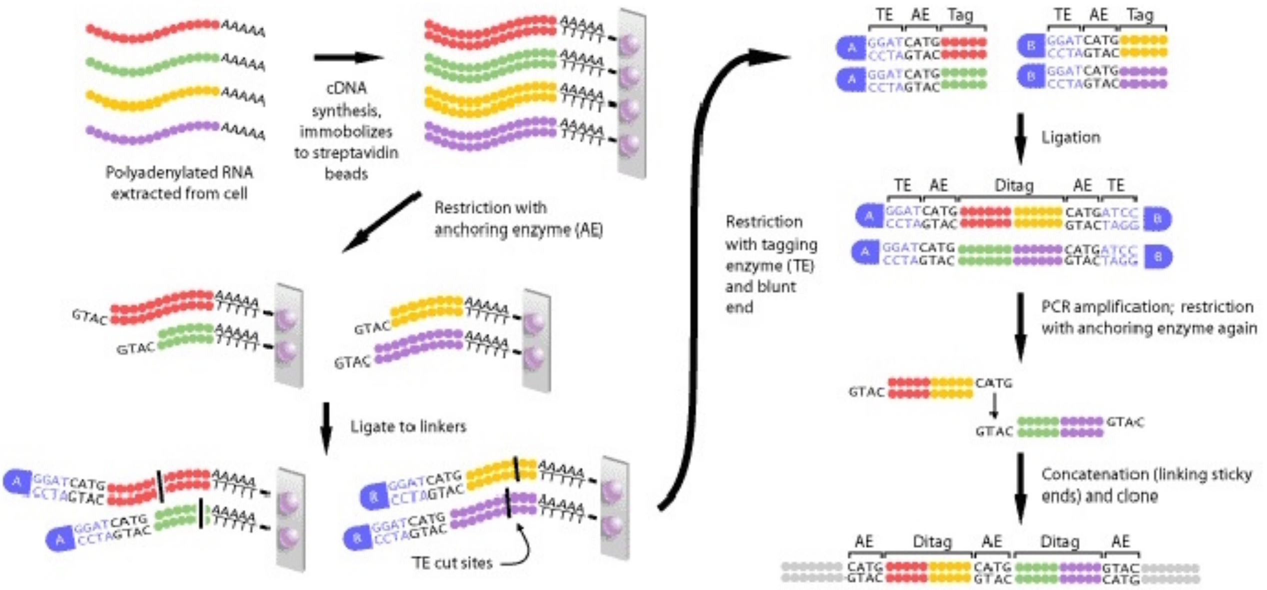

Polyadenylated RNA extracted from cell
cDNA synthesis, immobolizes to streptavidin beads
Restriction with anchoring enzyme (AE)
Ligate to linkers
TE cut sites
Restriction with tagging enzyme (TE) and blunt end
TE AE Tag
Ligation
TE AE Ditag AE TE
PCR amplification; restriction with anchoring enzyme again
Concatenation (linking sticky ends) and clone
AE Ditag AE Ditag AE
GGATCATG
CCTAGTAC
GTAC
CATG
AAAAA

Figure 13. Visual representation of the SAGE experimnet

It was not until the first high-throughput DNA sequencing methods were developed that sequencing a whole transcriptome seemed an overwhelming task: these new protocols permitted to sequence millions of bases in a very short time compared to traditional approaches and although they were hardly accessible when they came out, they grew in popularity and the per-base sequencing cost has decreased ever since.

RNA Seq

By replacing purification of DNA with RNA and retrotranscribing RNA to cDNA ("Central dogma reversed," 1970), the NGS methods could be used to inspect RNA expression. This approach rose to popularity soon after the commercialization of the first NGS sequencers and has been known as RNA-Seq (Costa, Angelini, De Feis, & Ciccodicola, 2010; Tang et al., 2009; Wang, Gerstein, & Snyder, 2009). Early applications of this technique date back to 2007, when Emrich et al. combined a laser micro dissection performed on maize with a run of 454 sequencing, producing ~260,000 reads from purified RNA (Emrich, Barbazuk, Li, & Schnable, 2007). This work allowed the annotation of many maize genes as well as showing the potential of NGS transcriptomics. In the same year, the same team used 454 for the discovery single nucleotides polymorphisms (Barbazuk, Emrich, Chen, Li, & Schnable, 2007) and since then the number of paper published on NGS transcriptomics has grown exponentially.

Typical workflow for RNA-Seq experiments

1. *RNA extraction and enrichment*

A typical RNA experiment starts with the isolation and purification of the total or part of the RNA present in the cell. The most common approach to isolate mRNA from the bulk of RNAs is known as poly-A tailing: bearing in mind that mRNA has a poly-adenylated tail, it is easy to extract it by hybridizing it on poly-T oligo covered magnetic beads. Since most of the RNA in the cell (more than 90%) consists of ribosomal RNA (rRNA) (de Leeuw, Slagboom, & Vijg, 1989), most of the current RNA isolation protocols rely on rRNA depletion: high levels of rRNA in the sample result in a high signal-to-noise ratio that can make detection of the RNA of interest difficult. As an example, Lifesciences Ribominus™ (Cui et al., 2010) technology utilizes specific locked nucleic acids to bind rRNA ribosome binding sites (Kozak, 1987; Shine & Dalgarno, 1975) and subsequently remove them from the total population of RNA with the aid of streptavidin-coated magnetic beads.

2. *RNA fragmentation*

After purification and isolation, the RNA needs to be sheared (Quail, 2001) into fragments in the range of 100-300 nucleotides each. This can be achieved via different methods, the most common being sonication (Sambrook & Russell, 2006b), nebulization (Sambrook & Russell, 2006a) or hydrolysis. These methods can be utilized on both single strand RNA or double strand reverse-transcript cDNA (later described in step 3). While RNA fragmentation provides a better coverage, it suffers form both 5' and 3' ends depletion (Wang et al., 2009). On the other hand, reverse-transcription performed with poly dT-oligomers, which bind to the 3' poly-A tails, is strongly biased towards 3' end of transcripts instead, and reverse-transcription with random hexamers results in an under-representation of 3' ends (Wang et al., 2009). This is from the reduced number of priming positions at which the reverse transcriptase enzyme can start cDNA synthesis. Enzymatic methods are also available, but enzymes tend to have much stronger sequence specific biases cleaving than mechanical or chemical methods and are therefore less favored (Poptsova et al., 2014).

3. *cDNA synthesis*

The fragments of RNA (or complete RNA if the fragmentation step is performed after retrotranscription) are converted to cDNA: a double stranded DNA originating from the RNA, using the reverse transcriptase enzyme (Myers & Gelfand, 1991). This can be achieved in two ways:

A) Using short T's sequences (oligo dT) whih hybridize to the poly-A tail and function as primers for the reverse transcriptase enzyme was one of the first methods used, but the resulting cDNA was 3' biased, diminishing the 5' end sequence representation (Myers & Gelfand, 1991);
B) Random sequence oligos (generally examers), which hybridize to random sites on the RNA ensure an even distribution of the retro transcribed fragments, and do not suffer from ends biases. This method is the preferred one to avoid oligo dT problems (Stangegaard, Dufva, & Dufva, 2006).

Once the first strand of cDNA has been synthetized, the reverse transcriptase removes the RNA and the second strand of DNA is synthetized by DNA pol I and ribonuclease H. cDNA has the advantage of being much more stable than single strand RNA, and can be easily amplified via PCR (Mullis et al., 1986).

4. *Adapters ligation*

In this step, the sticky ends resulting from cDNA synthesis are cleaved, resulting in blunt end double strand cDNA. A single adenine nucleotide is added at 3' ends, to facilitate the ligation of the adapters (or linkers), showing a sticky T nucleotide, which contain an index (also called barcode sequence) that is useful to keep track of the experiments once the total cDNA/RNA is sequenced (Chen et al., 2012). Once both adapters have been added, the cDNA is almost ready to undergo sequencing.

5. *Size selection*

To ensure that all the molecules have the same lengths, a gel electrophoresis of the total cDNA is performed (Aaij & Borst, 1972). The band corresponding to the expected length of the fragments (plus adapters) is extracted from the gel and the rest discarded. This step also is useful to eliminate un-ligated adapters and sequence artifacts, such as adapter-

adapter molecules which contain no cDNA fragments. Also, the sequencing of same-length fragments ensures that the read length resulting from sequencing is the same, effectively avoiding sequencing errors.

6. *Amplification and generation of DNA clusters*

As shown in the chapter regarding the NGS technologies, all the methods make use of PCR (Mullis et al., 1986) to amplify the starting population of cDNA fragments (Illumina uses bridge amplification for example). The immobilized spots on the supports are actually clusters of the same sequence repeated over and over, and ensure that the detection of the base incorporation is strong enough to be correctly interpreted by optical sensors.
After step 6, the NGS method of choice starts its workflow to determine the sequences of all the cDNA fragments. This is basically the common approach to RNA samples preparation for RNA-Seq experiments.

Reads Quality and experiment quality

As previously mentioned, one of the determinant factors in obtaining good results from the RNA-Seq experiments is to examine and evaluate the library quality and the sequencing performance. In a NGS experiment, low quality starting material inevitably lead to reading/processing faults by the NGS sequencer, that is, base miscalling. To avoid problems arising from these reading faults, it is common procedure to eliminate reads with low base quality confidence: information on base calling confidence is extracted from the variation of Phred values in the sequence (Ewing, Hillier, Wendl, & Green, 1998; Richterich, 1998). Phred quality scores were firstly introduced in Phrap program (Ewing & Green, 1998; Ewing et al., 1998), used in the Human Genome Project to automate sequencing. To determine a sequence score, phred algorhytm takes into account many factors, such as signal intensity and its relative peak shape, for each detected nucleotide incorporation. It then proceeds to compare the data to previously obtained data on the same type of nucleotides, stored in lookup tables, to assign a quality score to the sequenced nucleotide. Different tables are used according to the type of machine used and the chemical methods used by the same machine. Given the accuracy of the method, many manufacturers currently use phred-like scores in their sequencers.

Phred quality scores (Q) are defined as a property which is logarhytmically related to the base calling error probability (P):

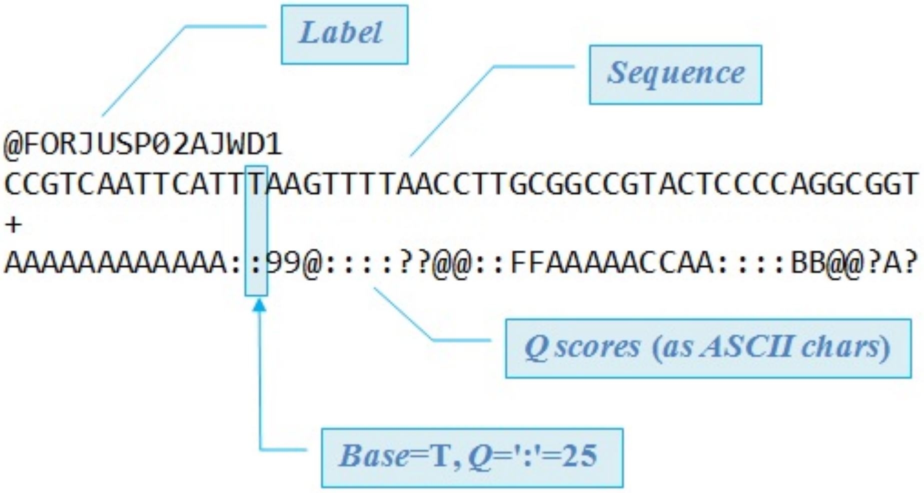
$$Q = -10 \log_{10} P$$

As an example, a Q value of 10 translates to a chance of calling the wrong base of 1 in 10, that is, a confidence of 90%. A Q score of 50 has a chance of miscalling of 1 in 100,000 and so on, as shown in the table. There is no “exact rule” to tell if the quality of a particular sequence is good or bad, as this depends on the purpose of the study. For example, an expression level analysis requires less read quality than investigating for SNPs or transcript variants. Generally, a score over 30 is considered very good, between 20-30 good and scores below 20 indicate poor quality (X. Li, Nair, Wang, & Wang, 2015).

Usually, quality scores and fragments sequences are stored together in FASTQ format (Cock, Fields, Goto, Heuer, & Rice, 2010). A brief overview of FASTQ format is given in figure. Due to the nature of scores being numbers, common compressing methods compress the value in ASCII format, which saves disk space and its length is the same of the sequence. Phred scores are reported from 33 to 126 (ASCII “!” to “V”), but SOLiD still uses numerical values to indicate base quality (Castellana, Romani, Valente, & Mazza, 2013).

Another way to determine reads quality is GC content (guanine-cytosine content). GC content represent the percentage of bases in a DNA sequence that are either G or C and is one of the most simple way to measure nucleotide composition of a DNA sequence. The reason behind GC choice over AT is that GC has a more direct biological meaning: not only the GC pair bond is stronger than AT (3 hydrogen bonds versus 2) but in the PCR amplification processes the GC percentage is critical in predicting annealing temperature (Yakovchuk, Protozanova, & Frank-Kamenetskii, 2006) and ensure better amplification results and it is known that information encoding sequences on the genome are richer than “junk DNA” in cytosine and guanine (Sumner, de la Torre, & Stuppia, 1993).

Therefore, GC content evaluation represents a simple yet critical step in transcription detection and abundance quantification.



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Figure 14. FASTQ read in Illumina format. Normally, there are 4 lines per read: the firts line starts with a '@', followed by a unique label; the second contains the read sequence; the third starts almost always with a '+', sometimes contains a copy of the label; the fourth contains quality scores in ASCII format. Quality score is referred to the nucleotide in the same column.

Transcriptome Assembly

Ab initio transcriptome characterization is based on de novo assembly of transcriptomes (Martin & Wang, 2011; Yassour et al., 2009). This step is essential when a reference gene annotation is not available, and anyway useful for the identification and investigation of novel alternative splicing events. Recently developed methods make use of de Brujin graphs (De Bruijn, 1946) and employ overlapping k-mers to assemble short reads into contigs. However, both sequencing errors and the occurrence of alternative splicing events in eukaryotes make the process of transcritpome assembly more challenging, both in terms of analysis and computational complexity (Lin et al., 2011).

Nevertheless, reference genomes/transcriptomes are available for only a small fraction of organisms, and since genome assembly is a complex and resource draining operation, transcriptome assembly has become a less costly and effective way to study organisms, facilitating read mapping, phylogenetics, marker constructions and other downstream applications. Several tools are now available for *de novo* assembly of RNA-Seq like Trans-

ABySS (Robertson et al., 2010), Velvet-Oases (Schulz, Zerbino, Vingron, & Birney, 2012), and SOAPdenovo-trans (Xie et al., 2014).

Reads alignment

When a reference genome sequence is available, the first step is the reassignment of the DNA fragments extracted to their original location on the genome: this operation is called *read mapping*. In other words, each read produced has to be matched (aligned) to the genome (or, similarlym transcriptome) in order to recover its original position. Although in theory this would correspond to finding the exact matching sequence, for which optimal solutions already exist, in practice only a small fraction of reads map without errors on the reference sequence. This is mainly because of two reasons:

1) *Nucleotide polymorphisms*: albeit having a reference genome on which to align reads, every organism slightly differs from it, notably in single nucleotide mutations scattered all over the genome. These mutations are called SNPs (single nucleotide polymorphisms). Thus it is possible that a read does not properly align to reference sequence due to the presence of one or more of these point mutations (Nielsen, Paul, Albrechtsen, & Song, 2011);
2) *Sequencing errors*: as previously mentioned multiple times, sequencing errors are totally normal in a NGS experiment and are caused by either human errors in preparation of the samples, incorrect base calling and enzymatic errors in replication/amplification cycles (Wall et al., 2014).

Due to the large throughput of NGS methods, and the need to process millions of short sequences by mapping them on the reference genome, traditional aligning algorithms are not suited anymore for the task. Classic alignment methods such as Smith-Waterman dynamic programming (T. F. Smith & Waterman, 1981), or indexing of k-mers in the template sequences (BLAT) (Altschul, Gish, Miller, Myers, & Lipman, 1990) or a combination of the two approaches (BLAST, BLAST+) (Camacho et al., 2009), while efficient in searching through databases to find homologue sequences, are not suited to the alignment of very large number of short reads (especially in the case of Illumina) to a reference sequences in the order of several Gbs, and do not take into account error rates and types specific of the sequencing platform. To overcome the need of expensive dedicated computing hardware for the alignment of reads, the development of fast, yet accurate aligners (using heuristics) has boomed in the last years, and new methods are published on a practically weekly basis.

The problem at the base of read mapping is thus approximate pattern matching, in which usually up to two-three substitutions are allowed on tags in the range of 40-100 bp. Also, the presence of indels in reads or in the genome/transcriptome induced by the presence of same-nucleotide repeat sequences may cause problems in the sequencing process and must be taken into account when trying to correctly align reads.

Another issue that arise is that while a 30-40 bp should be long enough to univocally map onto a reference sequence (e.g. chances of randomly finding a 30bp sequence are $1/4^{30}$), often a read can match different positions. Thus, the problem can be formalized in different ways: from approximate pattern matching allowing point mutation to approximate pattern matching allowing indels. While the mostly used alignment algorithm and approaches to the problem differ, they are substantially based on the same principles: the indexing of the reference genome or reads (H. Li & Homer, 2010). Indexing allows for faster sequence accession and speeds up sequence match research. Although a genome is usually in the range of billions of bases and a normal NGS experiment output is measured in millions/hundreds of millions bases, indexing the genome for reads alignment, while costly in terms of memory usage allows for faster read mapping and is therefore the most used methods to develop aligners. On the other hand, indexing reads results in lower memory demands, but longer mapping times. Various indexing methods have been proposed; additionally, newer methods allow the presence of indels in alignments and implements several tricks to speed up reads mapping. As an example, base-calling quality scores can be used to make solution search spaces smaller, by limiting the mismatch tolerance only to bases with low quality scores that are therefore less reliable (A. D. Smith, Xuan, & Zhang, 2008). Moreover, as mentioned before, 3' biases can be taken care of trimming 3' end nucleotides that usually present low confidence values (Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011).

By aligning reads extracted from a RNA-Seq experiment on the reference genome three main problems arise, namely, the assignment of reads corresponding to exons belonging to different splicing isoforms originating from the same gene; the mapping of reads that span exon-exon junctions; the mapping of reads which contain part of the poly-A tail.

The correct assignment of a read ambiguously mapping on two or more isoforms is a problem not completely solved yet. While reads assignment can be easily done for genes that do not have isoforms, it might be difficult to achieve when more isoforms are present due to alternative promoters or alternative splicing events: a read can be assigned explicitly to an isoform if it maps to a segment which uniquely belongs to it; read assignment is ambiguous if it maps to a segment shared by different isoforms. Different approaches can be used to address this issue. For example RSEM (B. Li & Dewey, 2011) employs an Expectation-Maximization (EM) strategy to estimate the maximum likelihood of a read belonging to a given alternative transcript (B. Li & Dewey, 2011). Cufflinks (Trapnell et al., 2010) estimates isoforms relative abundance using a statistcal model in which the probability of observing each read is a linear function of the abundance of the transcripts from which the read could have originated.

Earlier, the most common way to overcome the exon junction problem was to prepare a library (or an array) containing all the known exon-exon junctions and computationally predicted ones (Johnson et al., 2003). After a first run of mapping on the genome, unmapped reads were extracted and realigned on the exon junctions library. However, this method had the drawback of not being able to give information on previously non-annotated splicing events, which led to the implementation of so-called “spliced aligners”, programs that enable the mapping of exon-junctions spanning reads with large gaps in the alignment (Ameur, Wetterbom, Feuk, & Gyllensten, 2010). Currently one of the most widely used aligners of this kind is Tophat (Trapnell, Pachter, & Salzberg, 2009), which implements a “exon first” method: in the first step, reads are mapped to the reference genome using an un-spliced read approach; then, the unaligned reads are collected and split in shorter sequences, and regions containing a high coverage of reads mapped at the previous step are searched to identify possible splice connections. Conversely, many other mapping programs, such as GSNAP (Wu & Nacu, 2010), use “seed and extend” methods, which split the reads into shorter sequences to give rise to candidate region for alignement. As of today, “exon first” approaches are more used due to their flexibility and mapping times, although they rely on euristics to align spliced sequences, which lower exact mapping confidence.

When dealing with reads containing poly-A residues, the approach more commonly used is to identify all the reads that did not initially map to reference genome and either begin with repeated As ot Ts (cDNA) in series of at least 4-5 nucleotides. Trimming the repeated segments and remapping the shortened reads often result in correct alignement to the genome, and can be useful to identify new ununnotated polyadenilation sites (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013).

Estimating gene expression levels

When examining transcript levels in a RNA-Seq experiments, some factors must be taken into account: the number of reads aligning to a transcript reflects the relative abundance of the transcript in the starting library (more copies of the same sequence mean it will be sequenced more likely), but it is also a function of transcript length (longer sequences produce more fragments in the cDNA fragmentation step). Therefore, when analyzing gene expression levels, it is not sufficient to define transcript levels as the number of reads belonging to a transcript, but values must be normalized, to allow for comparisons across different experiments. in (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008) was presented a method to normalize data based on these assumptions and they called the measure RPKM (reads per kilobase of exon model per million mapped reads):

$$RPKM = \frac{total\ reads}{mapped\ reads\ (millions) \cdot exon\ length\ (KB)}$$

$$RPKM = \frac{10^6\,C}{NL/10^3}$$

Where: **C** = number of mappable reads that fall onto the gene’s exons; **N** = Total number of mappable reads in the experiment; **L** = The length of a transcript (sum of the length of its exons) in base pairs.

Later, this definition was extended to expected fragments per kilobase of transcript per million mapped read (FPKM), to take into account paired end reads (Trapnell et al., 2010). These measures are not the only ones that can be used: many new methods are continually proposed and published, based on different assumptions.For example, more recently, modified RPKM/FPKM values, that take the mappability of different transcript regions into account have been proposed and successfully used, like TPM in the RSEM package (B. Li & Dewey, 2011).

Sequencing depth is another factor to bear in mind when quantifying gene expression: a low sequencing depth means that rarely expressed genes are harder to screen and detect. Thus, the development of more sophisticated statistical analysis approaches has become one of the most important problems to deal with for RNA-Seq and NGS data processing.

Differential expression

A classic approach to analyzing transcriptome-based experiments is the comparative study of two or more experiments conducted in different conditions, such as disease vs. healthy or treated vs. non-treated biological samples. Primary goal of this approach (known as differential expression, DE) is to identify differential expressed genes, that is, selecting genes that present statistically significant transcriptional variability across the samples. Since the first appearance of hybridization-based methods (microarrays) for the analysis of DE, many statistical methods to compare enrichment of probes have been proposed and are still debated, but they are quite well established, due to how popular arrays were in the last decades. The introduction of NGS methods in the analysis of transcriptomes, while increasing output and overcoming the boundaries of limited dynamic range of detection typical of microarrays, posed new problems for the analysis of data to estimate DE gene levels. Due to the digital nature of NGS data, most of the statistical analysis methods for NGS based DE experiments were borrowed from SAGE analysis workflows; nevertheless, some of the methods used for microarray are still a viable choice to analyze NGS data. Early methods included Poisson distribution, binomial and normal distribution to approximate read counts.

Later, it was demonstrated that SAGE count is actually overdispersed and therefore more variation than that expected by the sample is present. To ensure correct counts and differential expressed genes levels accuracy, several new methods that took overdispersion into account were developed: the tw-test (Baggerly, Deng, Morris, & Aldaz, 2003) was one of the first statistical methods that allowed to estimate DE by using “between library variation”, based on beta-binomial distribution. The same group, one year later introduced logistic regression applied to overdispersion to allow the direct comparison of more than two samples (Baggerly, Deng, Morris, & Aldaz, 2004). Lu et al (Lu, Tomfohr, & Kepler, 2005) proposed a similar approach using log-linear models.

Nevertheless, early RNA-Seq DE methods did not take overdispersion into account and were based on Poisson distribution (Burden, Qureshi, & Wilson, 2014). When overdispersion was proved to be present in RNA-Seq as well, methods were adjusted accordingly. EdgeR (Robinson, McCarthy, & Smyth, 2010), one of the most common R packages used to investigate DE, implements a method based on negative binomial distribution, inspired by the work of Robinson and Smyth, apllicable to both RNA-Seq and SAGE. In 2010 this method was further expanded by the work of Robinson and Oshlack (Robinson & Oshlack, 2010), which proposed a normalization approach to lower the bias caused by genes highly expressed in just one sample that could lower the detection accuracy of other genes expression. Another popular R package called DESeq (Anders & Huber, 2010) implements a similar method based on negative binomial distribution, but uses a posterior empirical Bayesian approach to investigate different DE models.

Main Results

Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity

Mitochondria

From ancient greek μίτος (*mítos*, “thread”) + χονδρίον (*khondríon*), diminutive of χόνδρος (*khóndros*, “grain, morsel”), the mitochondrion is often referred to as the “powerhouse of the cell” (D'Erchia et al., 2015), since it is the organelle in which cellular respiration takes place and produces most of the cell’s ATP (adenosine tri phosphate) supply. Apart from cellular energy generation, mitochondria are involved in several other cell processes (such as differentiation, apoptosis etc.) as well as playing a key role in the control of cell cycle and cell growth (Pesole et al., 2012). In eukaryotes, they are uniparentally inherited from the mother (Henze & Martin, 2003).

The origin of mitochondria are still debated, and two main hypotheses have arisen: the endosymbiosis hypotheses states that mitochondria were originally prokaryotic cells, which were able to process oxygen in a way that could not be achieved by eukaryotic cells and started living in endosymbiosis in the cellular matrix (Andersson, Karlberg, Canback, & Kurland, 2003); the autogenous one instead claims that mithochondria were initially part of the nucleus DNA and they were born during the divergence of eukaryotes from prokaryotes (approximately 1.5-2 billion years ago) as a stand alone organelle. The DNA travelled across the nucleus membrane, and formed a protein-proof membrane to en velop the DNA. Since the mithochondrion possesses a circular DNA chromosome and many other features in common with bacteria, the most accredited theory is the endosymbiosis one.

As mitochondrial DNA (mtDNA) is a relatively small, abundant and easy to isolate, it has been the favourite target of early genome sequencing projects (Borst & Grivell, 1978) and the nucleotide sequence of mtDNA of thousands of species has now been determined. Structure and gene organization of mtDNA is highly conserved among mammals

(Taanman, 1999). The chromosome is ~16,500 bp and encodes for 37 genes: redox proteins coding genes indispensable for the respiratory chain (subunits of respiratory complexes I, III, IV and V), ribosomal RNAs, and 22 tRNAs for protein formation, same found in prokaryotes. A single mitochondrion can contain up to 10 copies of the chromosome. Most information is encoded on the heavy (H) strand, with genes for two rRNAs, 14 tRNAs, and 12 polypetides. The light (L) strand encodes for eight tRNAs and a single polypeptide (Taanman, 1999).

Mammalian mtDNA is extremely organized, when compared to the rest of genome. Genes are intron-less, except for one regulatory region, and intergenic sequences are practically absent or limited to a few bases. As in prokaryotes, the mitochondrial DNA has a high coding DNA/non coding DNA ratio, absence of repeated genes. The rRNAs and

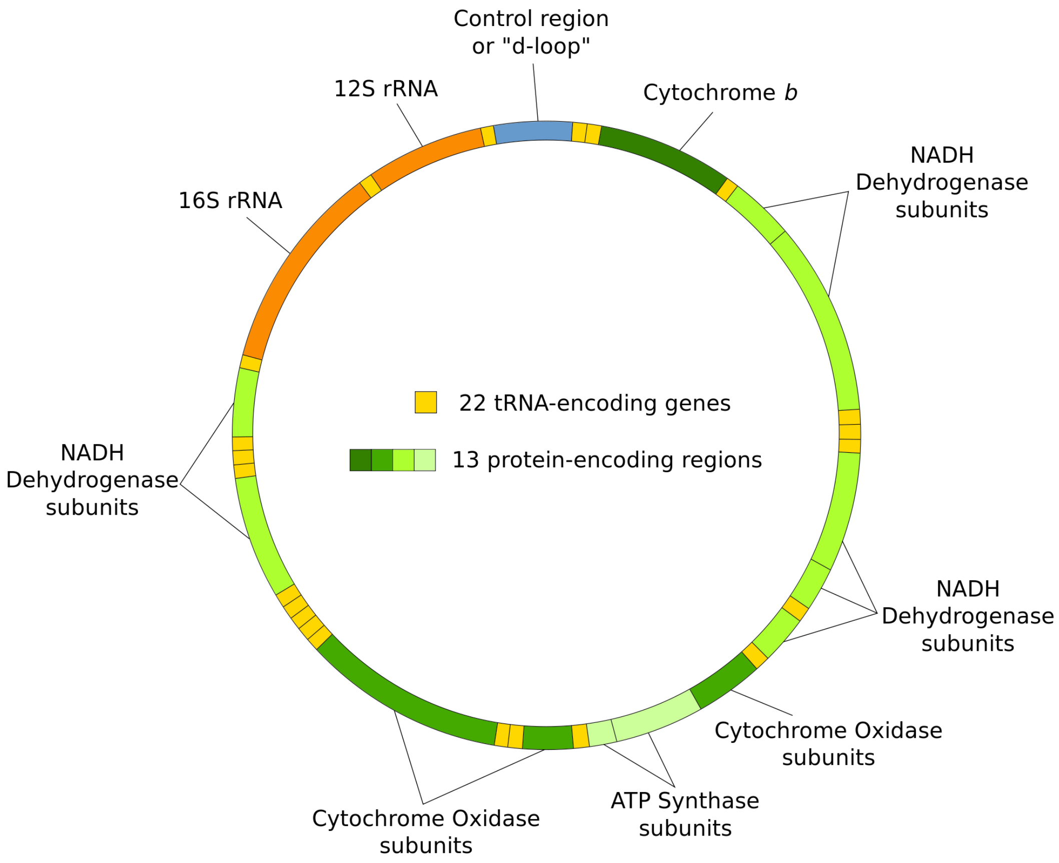


Figure 15. The mitochondrial DNA is a circular DNA molecule with a length of about 16500 bp. It encodes for a total of 37 genes, divided in 2 ribosomal RNA coding genes, 22 tRNA coding genes and 13 protein coding genes.

tRNAs originating from mtDNA are unusually small. Some protein coding genes overlap each other, and do not have complete termination codons, which are added during the process of poly-adenylation. In vertebrate cells that are metabolically active, a large part of the mtDNA contains a short three-stranded structure known as displacement loop or D-loop, in which a short DNA fragment complementary to the L strand displaces the H strand (Fernandez-Silva, Enriquez, & Montoya, 2003). This region is extremely conserved and functions as the major control site for mtDNA expression, containing the origin of replication and the major promoters for transcription. Once inititated at the D-loop, the L strand is transcribed as single polycistronic precursor RNA encompassing most of the genetic information contained in the strand. Subsequently, the precursor RNA is cleaved and polyadenilated to generate the mRNAs and the other RNAs. Most of the proteins needed by the mitochondrion to activate the respiratory chain are encoded in the mitochondrial proteins, but many key genes are localized in the cell nucleus, and they must be transported into the organelle, further avvalorating the endosymbiotic origin (Rackham et al., 2011; Scarpulla, 1997).

The mtDNA copy number per diploid nuclear genome correlates with ATP production and can range between 1000 and 5000 copies (Cavelier, Johannisson, & Gyllensten, 2000). This number varies dramatically among different tissues, cell developmental stages and is basically a reflection of the cell energy supply requirements. Altered copy number results in oxidative stress and linked pathological conditions, especially in tissues with high bioenergetics demands, such as muscles. It is thus not a surprise that mitochondria play a key role in the insurgency of hearth diseases and malfunctions (Crow, Mani, Nam, & Kitsis, 2004; Gustafsson & Gottlieb, 2008).

Next generation sequencing and mtDNA/RNA

It has been proved that off targets DNA sequences coming from whole exon sequencing experiments (WES) can be used to correctly assemble mitochondrial genome (Picardi & Pesole, 2012). It thus interesting to investigate if the data obtained from WES experiments can be used to infer the effective mtDNA copy number and whether or not the latter correlates with mitochondrial genes expression levels, measured by RNA-Seq.

The starting material of the study consisted of six different post-mortem human snap-frozen tissues (namely brain, heart, kidney, liver, lung and muscle) extracted from three unrelated healthy Caucasian individuals (males, respectively 47, 48 and 54 years old at the time of death) obtained from Cureline (South San Francisco, CA, USA). The three individuals are labeled as S7/11, S12/12 and S13/12 as summarized in table 3. DNA has been extracted from the samples and purified with DNeasy Blood and Tissue kit (Quiagen, 2015a) and quantitavely measured on NanoDrop 2000c (Thermo Fisher USA) (ThermoFisher, 2015) to check the optical density of the DNA sample. Total RNA was purified with the aid of RNeasy Plus Mini Kit (Quiagen, 2015b) and quality checked with Agilent Bioanalyzer 2100 (Agilent, 2015), with a resulting RIN (RNA integrity number) (Schroeder et al., 2006) in the range 5-7, acceptable for post mortem RNA extractions.

	Age	Sex	Race	Cause of Death	PMI
S7/11	47 Years	Male	Causcasian	Acute coronary syndrome	3
S12/12	54 Years	Male	Causcasian	Car accident	1
S13/12	48 Years	Male	Causcasian	Traumatic asphyxia	1

Table 3. Summary of the three individuals (labelled S7/11, S12/12 and S13/12) from which the samples were extracted. For each of them the respective age at the time of death, sex, race, cause of death and number of hours in which he samples were collected after death are reported (PMI).

For each tissue, a strand oriented (directional sequencing) paired-end library was prepared in order to keep information about which strand of the DNA the transcript was originated from. Strand orientation allows for detection of antisense transcription and can be useful to investigate gene regulation. The rRNA was removed from total RNA population and sequenced on the Illumina HiSeq 2000 platform (Illumina, 2015), generating 27 to 35 million reads per tissue sample. Each read length was 100 bp and insert size ranged from approximately 100 to 400 bp. This approach automatically excludes ncRNA from detection: the reference mitochondrial annotation we used contained the 22 tRNA located on the mtDNA, but as expected no read could be assigned to them due to absence of enrichment for transcripts <200 nucleotides long (such as miRNA and tRNA).

Results

Sequenced reads were initially mapped to reference human genome (assembly hg19) using Tophat (Trapnell et al., 2009) allowing at most two mismatches. An average of approximately 90% of total reads were aligned in correct forward-reverse strand fashion and with a maximum insert size of 1000 bp. Globally, 4~27% of total pairs were assigned to mtDNA, according to tissue sample. A further 8% of total reads could mapped also to Numt regions on nuclear DNA, either as a singleton (just one of the two reads in the pair correctly maps), chimeric pairs (reads in pair mapping on different chromosomes). To obtain a correct estimation of mitochondrion-originating reads, we decided to further investigate ambiguously mapped reads,. We discovered that for approximately 90% of the read pairs mapping on both mitochondrial and nuclear DNA, only one read mapped on nuclear DNA (mostly on CO1 or CO3 genes). Of all the correctly forward-reverse mapped reads, only 1% mapped on nuclear DNA too, but generally with a higher mismatch rate than in the mapping on mtDNA. This led us to the conclusion that correct pairs mapping on mtDNA can be safely used to estimate mithocondrial transcript levels. FPKM values were then obtained starting from read counts for all the mitochondrial annotated genes.

We found that overall, in all tissue samples, transcription levels of genes annotated on the forward and reverse strand of the mitochondrial chromosome were extremely unbalanced, with about one thousand fold increase for those mapping on the forward strand, while approximately 0.36% of the reads mapping on the forward strand resulted to be originated from the precursor RNA. On the other hand, 85% of the reads mapping to the reverse strand were found to be originated from precursor RNA. The largest fraction of total reads (~95%) originated from rRNA (12S and 16S) in all analyzed tissues, while other reads mostly mapped on protein coding genes, albeit with some enrichment on antisense genes (such as genes coding for 16S rRNA, CO1 and ND5). Antisense transcription was detected in proximity of ND5, ND6 and CytB genes, which has been reported as transcription of lncRNAs (Rackham et al., 2011).

Expression levels of the 11 protein coding genes localized on mtDNA showed high variability among tissue samples, reflecting the relative concentration of mitochondrial DNA in the different tissues. In contrast, as shown in Figure 16, relative expression levels of protein coding genes are practically constant within the same sample, due to different levels of steady-state expression of mature transcripts, a probable consequence of the various post transcriptional modifications that occur in mitochondria (Rorbach & Minczuk, 2012), conserved across all the tissues studied.

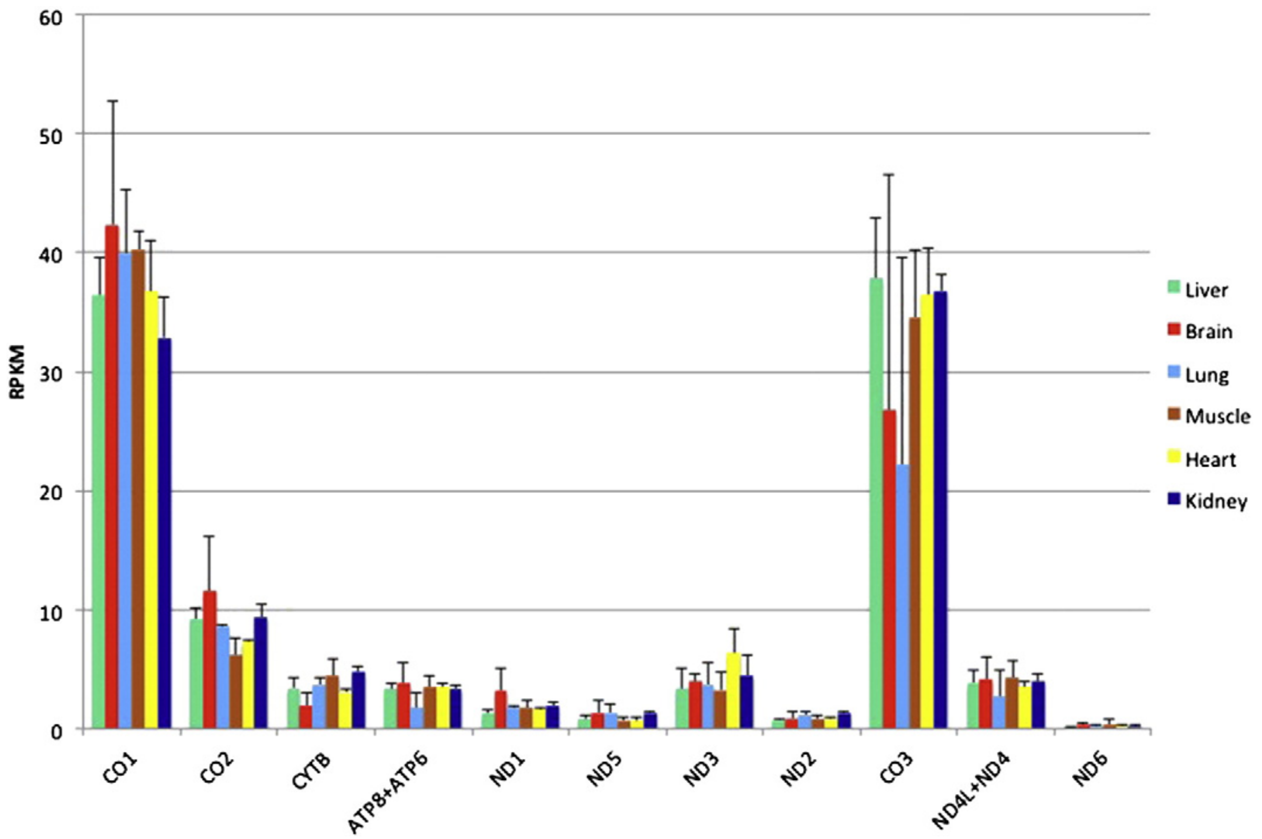


Figure 16. Expression levels of protein coding genes are practically constant within the same tissue samples.

We then performed a correlation analysis between mtDNA copy number (measured by qPCR analysis) and expression of genes localized on mtDNA, defined as the sum of FPKM values of non-ribosomal mitochondrial genes in the three different individuals. The result was a significant correlation (Pearson correlation of 0.81). Similar results were obtained by also considering mitochondrial rRNA-coding genes expression levels, or in general the fraction of total mitochondrial RNA in the sample (Pearson correlation value ~0.8).

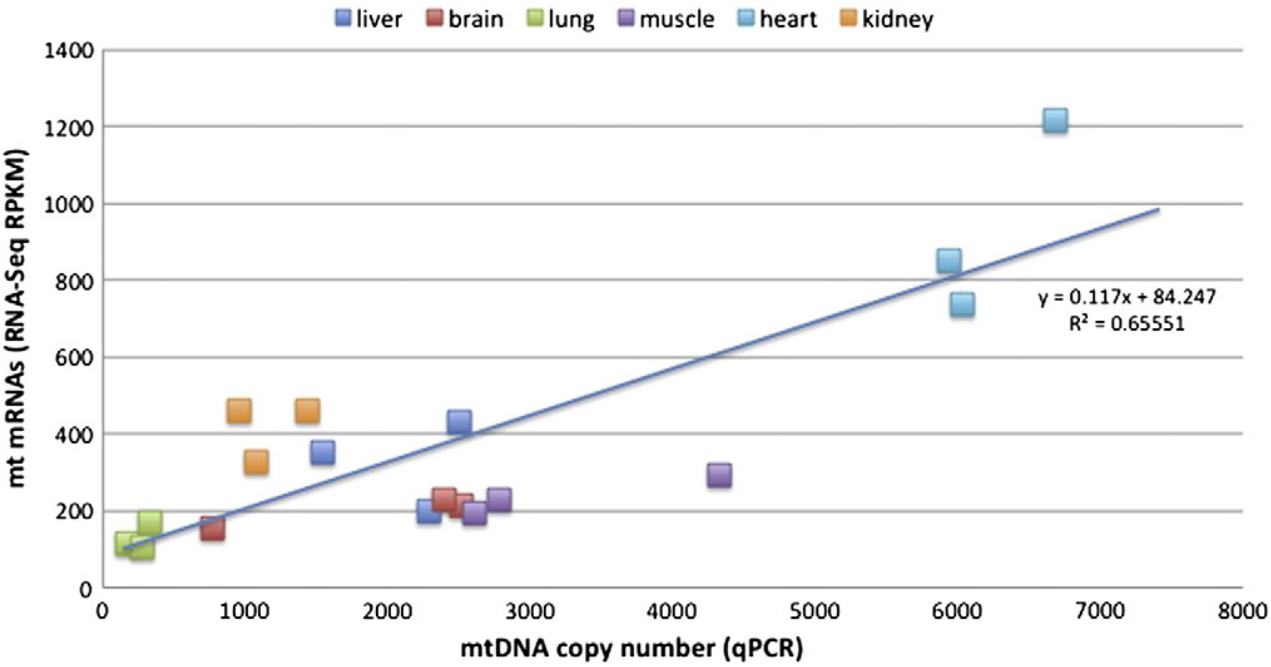


Figure 17. Correlation plot between mtDNA copy number measured by qPCR and mt mRNAs measured by RNA-Seq

By measuring relative mtDNA copy number in the six tissues, we found that heart samples held the highest mtDNA copy number, followed by skeletal muscle, brain, liver, kidney and lung. This result was consistent with the hypothesis that mtDNA copy number should be higher in tissues with high embolic and bioenergetics demands (e.g. muscles). In particular, the variation in relative mtDNA content was especially evident between muscle and brain, and the estimated low copy number in brain could be also attributed to the cause of death of individual S13/12, that is, traumatic asphyxia.

Quantification of mtDNA with WES off target reads proved to be an accurate method and its results correlated with qPCR mesurements of mtDNA levels with high statistical significance (bivariate linear correlation r^2= 0.92, P< 0.0001).

An entropy based framework for the identification of sample specific expression patterns and splicing events

The general method we introduced works on transcript levels derived from a series of RNA-Seq experiments. Samples can come from different tissues, from the same cell line in different stages or conditions, from different individuals, and so on. Each sample can be sequenced once or, better, in any number of biological replicates. The framework implements different statistical tests that, starting from the estimated expression levels of each RNA of the transcriptome investigated in every replicate, are able to answer the following questions:

- Which genes, or transcripts, show a biased expression pattern, that is, are significantly over- or under-expressed in one or more of the samples? For example, if samples correspond to different tissues, it detects tissue-specific genes and transcripts; if RNAs come from different time points, it detects which ones are significantly over- or under-represented at some time points, and so on.
- Which genes show a significant "isoform switch" across two or more of the samples? That is, genes that have two or more alternative transcripts changing in a significant way their relative abundance across the samples investigated. These genes might or might not be specific for some of the samples, since the relative abundance of the different transcripts might compensate each other making the overall expression of the gene uniform.
- The two above points can be studied also at the splicing event level: which splicing events (e.g. inclusion of a cassette exon) or alternative promoters are significantly over- or under-represented in one or more of the samples? Which present a significant "switch" across two or more of the transcripts?

Quite obviously, there is no unique definition of tissue- or in general "sample specific" genes, which can change significantly according to the criteria used and the case studies investigated. For example, one could require the expression of the gene in a single sample to be greater than *k* times its expression in any of the other samples. This would in turn single out genes specific for a single sample. Or, we could define a gene as sample specific if its expression in a sample is greater than *k* times the average expression of the gene across all samples. This second definition is less stringent, and could return genes specific for more than one of the samples analyzed. Approaches like the ones just mentioned, however, need the definition of explicit thresholds, that is, requiring that a gene has to have a *k*-fold increase of expression, with significantly different results according to the value of *k* chosen. Moreover, only the relative variation of expression across the samples is considered, and thus two genes with very different expression levels will be considered to be “equally significant” if they present the same fold change with respect to e.g. the average expression level.

In this work we first of all introduce a measure, based on information theory, assessing how much the expression of a gene across different samples differs from a uniform distribution. When applied to samples coming from different tissues, the latter characterizes for example genes that could be considered “housekeeping” (Eisenberg & Levanon, 2013) and, in general, without any significant change of expression in any of the samples. Starting from this measure, we derive a statistical test, returning the probability of the expression pattern observed to be actually resulting from a uniform distribution. In this way, the lower is the p-value is, the more distant is the distribution observed from the uniform one, and classical p-value thresholds of 0.01 or 0.05 for statistical testing can be applied, without the need of defining explicit thresholds for expression and/or fold enrichments.

After a subset of genes or transcripts has been singled out to have a distribution significantly different from the uniform case, further processing can identify in which of the samples lie the most significant differences, and thus which are the ones for which each of the transcripts is more “specific”. The same measures we introduce at the genes or transcript level can be applied also at the splicing or promoter level, thus identifying for example sample-specific splicing events or alternative promoter usage. Finally, we show how the same framework can be applied at the gene level, for the identification of genes that have the highest variability of isoform abundance across the different samples considered, comparing and summarizing the sample specificity of the alternative transcripts of the gene.

Detecting Sample Specific Expression Patterns

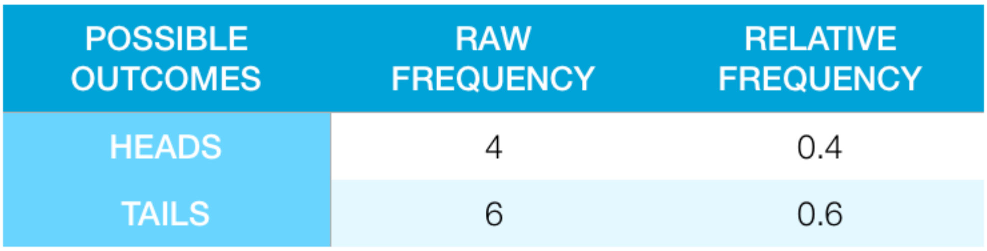
Given a set of gene expression measurements performed on *m* different samples or conditions, the problem of detecting “sample specific” genes can be informally defined as identifying those genes that present an expression pattern that seems to be biased towards one or more of the samples, that is, increase in a significant way their expression level in a limited subset of the samples. Viceversa, genes not showing any sample-specificity should present a uniform expression level across all the samples investigated.

The problem has been widely studied since the initial transcriptome studies based on technologies like microarrays. Among many others, an approach that was introduced for the problem comes from information theory and is based on *Shannon’s entropy* (from now on referred to simply as entropy). The entropy concept was introduced by Claude E. Shannon in the paper “A mathematical Theory of communication” in 1949 (Shannon, 1949). Entropy can be used to measure the uncertainty associated with a random variable, i.e., the expected value of the information in a message (bits in informatics). The idea behind entropy is that the less likely an event is, the more information it provides when it occurs. The probability distribution of the events, coupled with the information content of every event, yields a random variable whose expected value is equal to the average amount of information generated by the distribution. The general formula to compute entropy is:

$$H = -\sum_{i=1}^{m} f_i \log_b f_i$$

where f_i is the frequency with which the random variable is observed to assume the *i*-th value. Entropy comes in form of logarithm value, because of the additive property of logarithms. To understand this, let us consider this coin-flip example. A single coin flip provides a bit of information (the outcome can be either heads or tails with equal probability, therefore the two results can be described by using a single bit, 0 or 1) and m coin flips provide m bits of information. In general, an event that can lead to n equally possible outcomes needs $\log_2(n)$ bits to be represented. This rule holds true as long as all the outcomes are equally probable.

If an event occurs more frequently than others, it is less informative, and conversely rarely observed events hold more information: the observation of a rarer event has the effect of lowering the entropy value under the $\log_2(n)$ threshold, because the data will not be uniformly distributed. Entropy is 0 when only one single outcome is observed (minimum dispersion) and 1 when all the outcomes are equally distributed (maximum dispersion). As for the coin flip, consider the following table:



POSSIBLE OUTCOMES	RAW FREQUENCY	RELATIVE FREQUENCY
HEADS	4	0.4
TAILS	6	0.6

Table 4. Frequency of heads/tails in 10 coin flips.

The entropy of the observed distribution would be computed as follows: $H = -0.4 \log_2(0.4) - 0.6 \log_2(0.6) \approx -0.4(-1.3) - 0.6(-0.7) = 0.52 + 0.42 = 0.94$ bits

For our case, let T be a transcript (or a gene), and $t_1 \ldots t_m$ its expression levels in m different samples. In a RNA-Seq experiment, as discussed before these values are usually estimated with normalized FPKM or TPM values. The calculations we present can be however performed on any other similar normalized measure.

Let $\bar{t} = \sum_{i=1}^{m} t_i$ the overall expression value of T, and $f_i = t_i/\bar{t}$. The f_i values are the ones that can be used to compute *H(T),* that is, the entropy of the expression of the transcript across the *m* samples. Entropy *H(T)* will be minimum (zero) when the transcript is expressed only in a single sample. The more unbiased the expression is, the more *H(T)* will grow, and will be maximum if the transcript has identical expression values across all the samples. Suitable thresholds for *H(T)* can be then defined to detect genes with a “sample specific” expression pattern.

Nevertheless, this methods has two drawbacks. The first is that entropy calculation does not take into account the absolute expression levels, but makes use of relative variation across samples. In other words, a transcript with FPKM values of zero in all the samples but one, and a very low (<1) FPKM value in only one sample will have entropy zero, and result to be more “sample-specific” than transcript with a very high level (e.g. FPKM = 1000) in one sample, and significantly lower (e.g. FPKM = 1) in the others. The second drawback comes from the fact that the distribution of FPKM values across the samples is compared to background expected values of b_i = *1/m*. To take into account cases for which the distribution of background values is not uniform, relative entropy (or, the Kullback-Leibler divergence (Kullback, 1951) between the empirical distribution of the f_i values and theoretical one of b_i, such that $\sum_{i=1}^{m} b_i = 1$) can be employed:

$$RE(T) = \sum_{i=1}^{m} f_i \log_b \frac{f_i}{b_i}$$

The difference is that relative entropy is zero when $f_i = b_i$ for every *i*, and maximum where f_i = 1 for some *i*.

To take expression levels into account, we introduce a “weighted” version of entropy, that we will call *sample specificity index*, where each term is multiplied, instead for the respective frequency, for the corresponding expression value t_i:

$$SSI(T) = \sum_{i=1}^{m} t_i \log_b \frac{f_i}{b_i}$$

This value will be proportional to the value of a G-test (goodness of fit test, or log-likelihood ratio test), where the null hypothesis is that the observed values t_i result from a random distribution with expected frequencies b_i*:*

$$G(T) = 2\sum_{i=1}^{m} t_i \log_e \frac{f_i}{b_i} = 2SSI(T)$$

The *G(T)* values are distributed with a chi-square distribution with *m – 1* degrees of freedom. This fact allows us to associate a p-value with every *SSI(T)* value, by multiplying it by two and using *b* = *e* as the base of the logarithm (natural logarithm). The p-value will denote the probability of obtaining the t_i observed values by chance given a random background distribution with frequencies b_i = *1/m.* . We call this computation the *general sample specificity test.*

To assess the overall sample specificity of a single transcript we can assume a uniform background distribution with b_i = *1/m*, but the same framework can be used with any other assumption on the theoretical distribution of the expression values, as we show in the “Isoform Switch” section. Typical p-value thresholds of 0.01 or 0.05 can be employed to single out transcripts with significant sample-specific bias, but since thousands of statistical tests are performed, one for each annotated transcript, p-values need to be corrected for multiple testing. In this work we employed the Benjamini-Hockberg procedure (Benjamini & Hochberg, 1995), ranking the genes or transcripts according to increasing p-values, and given a significance p-value threshold of a we considered significant all transcripts for which $G(T) \leq \frac{k}{N}\alpha$, where N is the overall number of transcripts and *k* is the ranking of the gene or transcript.

Identifying Specific Samples

The SSI(T) method and its relative p-value are used to indicate if a transcript T shows an expression pattern that deviates significantly from a uniform one. The samples for which T will show higher specificity will be the ones contributing more to the sum in the SSI(T) formula. One solution is to arbitrarily choose a threshold and consider significant all the transcripts whose contributions exceed the threshold. In order to avoid the problems arising from defining an explicit threshold, we decided to opt for a different strategy: once a transcript T has been singled out to be “sample biased”, we compute the SSI(T) and the relative p-value again by comparing, for each sample, just two values, i.e. the expression value of T in the sample considered and the overall expression of T in the others. Hence, the expected values on m samples are now $(\frac{\bar{t}}{m})$ for the sample considered and $(m-1)\bar{t}/m$ for the others. We call this computation the *single sample specificity test.*

All in all, a transcript *T* that passed the general sample specificity test can be reported to be specific for those samples with expression value higher than the average and that in the pairwise test just described yield a p-value lower than a threshold α. Notice, in fact, that this test will yield as significant also those transcripts with expression values significantly lower than the average in some samples, for which they can be then considered to be “avoided”.

Detecting Sample Specific Isoform Switches

Let G be a gene, and T_1, …, T_n *n* different alternative transcripts (isoforms) for the gene, produced by alternative splicing, alternative poly-adenylation, promoter usage, and so on. Sample specific alternative transcripts for G can be detected by computing *SSI* values and the sample specificity tests for each T_i, and by determining whether one or more transcripts result to be sample specific for different samples.

One the most interesting evolution of the method would be to define a single index at whole gene level, in order to detect significant isoforms switches, that is, significant changes in the relative abundance of the isoforms across the samples analyzed. For example, consider a gene yielding two different isoforms, both showing equal level of expression (l) in all the samples, with the exception of two different samples in which either one is expressed al level 2l. It is very likely that none of the two transcripts will show a sample bias large enough to be considered sample-specific, but what is more “sample specific” in this particular case is instead the combination of the isoforms, that is, the pattern of on-off-transcription observed in the two samples.

The idea is to estimate the expected frequency of usage of each isoform from the data itself, and compare the actual observed frequency in each sample with the expected frequency derived from all the other samples. If we represent the expression level of a gene with *n* alternative transcripts across *m* samples with a *n* x *m* matrix, then sample specific transcripts were singled out by performing a G-test on the rows of the matrix, while sample specific isoform switches are detected by G-tests on the columns of the matrix.

Let *m* be the number of samples, and let *k* be the sample for which we want to compute the expected frequencies. Let t_{i1},…,t_{im} be the FPKM values for transcript T_i measured across the samples. Let $\bar{t}_i^k = \sum_{j \neq k}^{m} t_{ij}$ be the sum of the FPKM values for T_i across all the samples different from *k*. The overall frequency of usage for transcript T_i in all the samples different from *k* can be thus defined by dividing $\bar{t}_i^k$ by the overall expression value of all the transcripts of gene G in all the samples different from *k*:

$$f_i^k = \frac{\bar{t}_i^k}{\sum_{j=1}^{n} \bar{t}_j^k}$$

Likewise, we compute the sample frequency of usage for transcript T_i in each sample $k = 1,\ldots,m$, dividing its FPKM value by the overall FPKM of G in the sample:

$$f_{ik} = \frac{t_{ik}}{\sum_{j=1}^{n} t_{jk}}$$

Therefore, a gene with no significant isoform switch across the samples will have a uniform frequency of usage for each alternative transcript. In other words, for each T_i the f_{ik} values in the different samples will be close to the expected f_i^k value. Vice versa, let us consider a gene with two alternative transcripts, expressing one in half of the tissues and the other in the other half, at the same FPKM value. The overall frequency of usage will be ½ for both the transcripts, but in each sample there will be one with frequency 1 and the other with frequency 0. Starting from these values, we can define the *isoform switch score* of the gene in each sample *k*:

$$IS(G_k) = \sum_{j=1}^{n} t_{ik} \log_b \frac{f_{ik}}{f_i^k}$$

Once again, the *IS* values are distributed with a chi-square distribution, this time with *(n-1)* degrees of freedom where *n* is the number of alternative transcripts for gene *G*. Suitable threshold for p-values can be then set in order to detect tissues in which genes show a significant isoform switch with respect to the others. For the gene across all samples, we can also define $IS(G) = 2\prod_{k=1}^{m} IS(G_k)$, in order to rank genes according to the variability of isoform usage across all samples.

Identifying Sample Specific Splicing Events

The overall strategy and measures we just defined for detecting sample specific transcripts can be also applied at the whole gene level (by applying them to the sum of the expression values of all the alternative transcripts of the gene in each sample), or vice versa for each single alternative splicing event characterizing the gene. Since an alternative splicing event results in the choice on whether including or not a given fragment of the primary transcripts in the mature ones, the SSI and G-test can be computed for the expression values of all the transcripts that include a given fragment, in order to determine if the fragment is included in a “sample specific” way. For example, in case of a cassette exon, we sum the expression values of all the transcripts that include the exon as the overall “expression value” of the exon across the different samples, and then derive the SSI value and corresponding p-value for the exon starting from these values. This strategy permits to identify which alternative exons or splicings which are included in tissue specific transcripts.

However, we can also identify sample specific *switches* at the detail of single splicing events, by modifying accordingly the computation of the “isoform switch” score. For example, let *e* be a cassette exon of gene G and *k* one of the samples. We want to determine whether exon *e* is included in the transcripts produced in sample *k* with a frequency significantly different from the other samples. We first sum the FPKM values in all the samples but *k* of the transcripts that include *e*. Let $\bar{t}_{in}^{k}$ be this value. Then, we sum the values of all the transcripts of G in all the samples with the exception of *k*. Let $\bar{t}_{out}^{k}$ be this value. Then, the expected frequency of inclusion for exon *e* in sample *k* will be given by:

$$f(e,k) = \frac{\bar{t}_{in}^{k}}{\bar{t}_{out}^{k}}$$

Then, let $t_{in}(e,k)$ the sum of the expression values in sample k of all the transcripts that include exon *e*, and *G(k)* the overall expression value of gene G in sample *k*. We define then $t_{out}(e,k) = G(k) - t_{in}(e,k)$. Clearly, the expected expression value for the transcripts that include *e* will be *f(e,k)G(k)*, and *(1-f(e,k))G(k)* for those that do not include *e*. Therefore, we can compute a *exon specificity score* for e in sample *k*, denoting how much specific for sample *k* is its inclusion in mature transcripts, by comparing the observed values to the expected ones:

$$ESS(e,k) = t_{in}(k,e)\ln\frac{t_{in}(e,k)}{f(e,k)G(k)} + t_{out}(k,e)\ln\frac{t_{out}(e,k)}{\big(1-f(e,k)\big)G(k)}$$

Once again, the value *2EES(e,k)* follows a chi-square distribution, this time with one degree of freedom. Samples for which the corresponding p-values are below a given threshold (e.g. 0.05 or 0.01, corrected for multiple testing) are in turn those where there is a significant difference, or switch, in the inclusion/excision of the exon with respect the other samples.

Using replicate experiments

In RNA-Seq analyses, it is common practice nowadays to perform biological replicates, that is, sequence more than once each sample in order to obtain more robust results less affected by experimental noise. The significance measures we just introduced can easily accommodate this point, by simply treating each replicate as a separate sample by itself.

Consider for example, as in the results we present in the next section, an analysis performed on six samples. Suppose we have only one replicate per sample, and a transcript has FPKM value of 10 in one sample and 1 in all the others. Computing the sample specificity score will yield for the transcript a p-value of 0.002, since the ten-fold increase of expression in one sample is attenuated by the overall low expression. In turn, the transcript might or might not be considered significant according to the correction for multiple testing employed. But, if we had another replicate for each tissue, with values very close to the first one, the overall p-value for the transcript will be approximately 0.0002, that is, one order of magnitude lower. With three replicates, once again with very close values, the p-value will be $5x10^{-6}$, and so on. The overall effect is that the more replicates producing consistent FPKM values are available for each sample, the more robust the FPKM estimates are, and this in turn makes more significant the difference from the uniform background distribution, and all the different replicates for the sample(s) for which a transcript is specific will turn out to be significant.

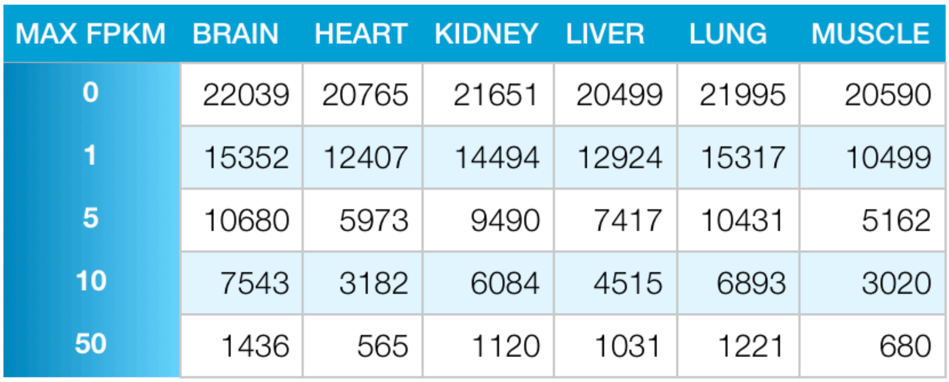
One advantage of the measures we introduced is thus that they can be applied by processing the data keeping the replicates separate, without the need to computing mean or standard deviation values. In other words, we can apply the specificity measures defined by considering each replicate as a separate "sample". Then, a transcript will be considered to be "sample specific" if the statistical tests consider it to be significant in all (or, a majority of) the replicates of a given sample. The only difference is that in the statistical tests determining whether a transcript can be considered to be specific for a given sample, the expected value will be computed without taking into account the expression values of the other replicates. Another useful side effect is that in this way transcript producing inconsistent measures can be singled out, that is, those transcripts or genes resulting to be specific only for a subset of the replicates of a given sample can be identified and excluded from further analyses since their variability can be considered too high.

Identification of tissue-specific isoforms reveals widespread individual-specific expression patterns

The same set of 54 samples, that is, derived from six tissues of three different individuals, employed in the mtDNA/RNA study was used to produce these results. For the quantification of expression levels each sample was processed the sequences using the Rsem software package (B. Li & Dewey, 2011), using the UCSC human gene annotation (hg19, version 2013-06-14) as a reference. We chose the latter because we consider it to be the one to provide the best compromise between reliability of the transcripts included (the RefSeq collection integrated with additional full length RNAs from Genbank) and alternative transcript abundance amongst the different alternatives available (RefSeq, ENSEMBL, GENCODE). The annotation comprises 78,289 transcripts assigned to 28,812 different genes.

We processed the data computing the general sample specificity index and corresponding p-values first on each gene (on the sum of the FPKM values of its transcripts), and then for each single transcript, keeping each replicate separate from the others as described in the previous section. That is, the test was performed over 54 samples.

At the gene level, by applying the general sample specificity test with a Benjamini-Hockberg corrected p-value threshold of 0.01, 11,476 genes of the 28,812 available in the annotation (about 40%) showed an expression pattern with a significant deviation from the uniform distribution. If restricted to the 23,681 genes with transcript level greater than zero in at least one sample, the percentage rose to almost half of the genes (FPKM > 1 in at least one sample: 18,674; FPKM > 5 in at least one sample: 14,365).



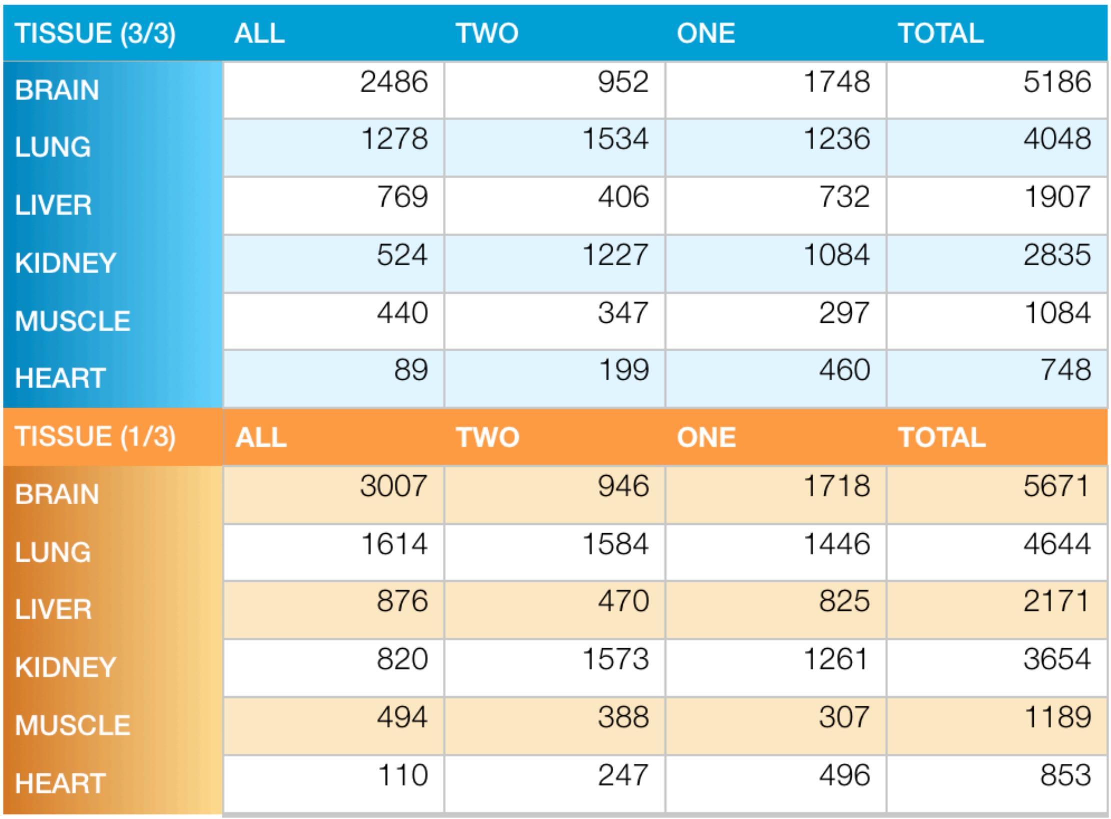
MAX FPKM	BRAIN	HEART	KIDNEY	LIVER	LUNG	MUSCLE
0	22039	20765	21651	20499	21995	20590
1	15352	12407	14494	12924	15317	10499
5	10680	5973	9490	7417	10431	5162
10	7543	3182	6084	4515	6893	3020
50	1436	565	1120	1031	1221	680

Table 5. Number of genes with Max FPKM > threshold in each tissue.

We deemed a gene that passed this initial test to be specific for a given individual in a given tissue if it passed the *single sample specificity test*, with an expression value higher than the mean across the samples, in *all* the three replicates for the sample itself with a p-value threshold of 0.05, and to be in general “tissue specific” if specific for the tissue in at least one individual. As shown in Table 6, the tissue with the highest number of genes showing a significant bias towards the tissue in at least one individual was brain, with more than 5000 genes, followed by lung (about 4000), kidney (3000), and liver (about 2000). Quite surprisingly, however, in each tissue only a fraction of the genes resulted to be significantly biased for the tissue *all* the three individuals, that is, when significantly over-expressed in a given tissue resulting to be so for all the three individuals and all the nine replicates available.

In other words, the expression of the gene within the same tissue showed changes among the three individuals of such a magnitude that according to our test the gene can be considered to be over-expressed in the tissue only in one or two of them. It can be seen also how the trend towards individual specificity changes according to the tissue, with brain having most of the tissue specific genes in turn specific in all the three individuals, while vice versa heart has the large majority of genes specific for only one individual.

In order to determine whether the criterion we adopted was too stringent, we considered a transcript to be individual specific in a given tissue if it passed the single sample specificity test in at least one replicate out of three. Individual-specific expression remained however a widespread phenomenon, with only marginal differences from results obtained with the more stringent criterion (Table 6).



TISSUE (3/3)	ALL	TWO	ONE	TOTAL
BRAIN	2486	952	1748	5186
LUNG	1278	1534	1236	4048
LIVER	769	406	732	1907
KIDNEY	524	1227	1084	2835
MUSCLE	440	347	297	1084
HEART	89	199	460	748

TISSUE (1/3)	ALL	TWO	ONE	TOTAL
BRAIN	3007	946	1718	5671
LUNG	1614	1584	1446	4644
LIVER	876	470	825	2171
KIDNEY	820	1573	1261	3654
MUSCLE	494	388	307	1189
HEART	110	247	496	853

Table 6. Number of tissue-specific genes in the six tissues detected in all three, two and one individual, in the case all three replicates (3/3, upper) or at least 1 replicate (1/3 lower) passed the specificity test.

Thus, a large fraction of genes showed an expression pattern significantly biased towards only one sample of only one individual, and in general about one third of the genes that could be considered to be specific for one or more of the tissues studied could be considered to be so in only one individual. Remarkably, with the first criterion (three replicates out of three significant) we obtained that 2007 genes were to be considered purely “individual specific”; that is, with expression significantly biased towards one or more tissues but in only one individual. Gene DDX3Y (Figure 18) is a typical example of the complex mix of individual- and tissue-specific expression emerging from the data, and how a gene can result to be tissue-specific for only one individual. It has an expression pattern clearly biased towards individual S13, to the point to be considered tissue specific in kidney, lung and brain only for S13, since the expression values in the latter have the effect of producing a higher mean value across the samples, and hence the gene fails to pass the test in the other two individuals. On the other hand, if only S7 and S12 had been considered, the gene would have resulted to be kidney and lung specific in both, but not brain specific. SNHG8 (Figure 19) is over-expressed in individual S7, and reported as “tissue-specific” in four tissues (kidney, liver, lung, and muscle). If S7 had not been included, the gene would have been “specific” for brain and not for muscle.

These examples clearly show how the “tissue-specificity” of a gene is a concept very far from being straightforward, and can be significantly influenced by the provenience of the samples studied, a factor often overlooked. Indeed, like in the two examples just shown, almost 300 genes out of 2007 had individual specificity as a key feature stronger than tissue specificity, that is, resulted to be specific for more than one tissue in only one individual.

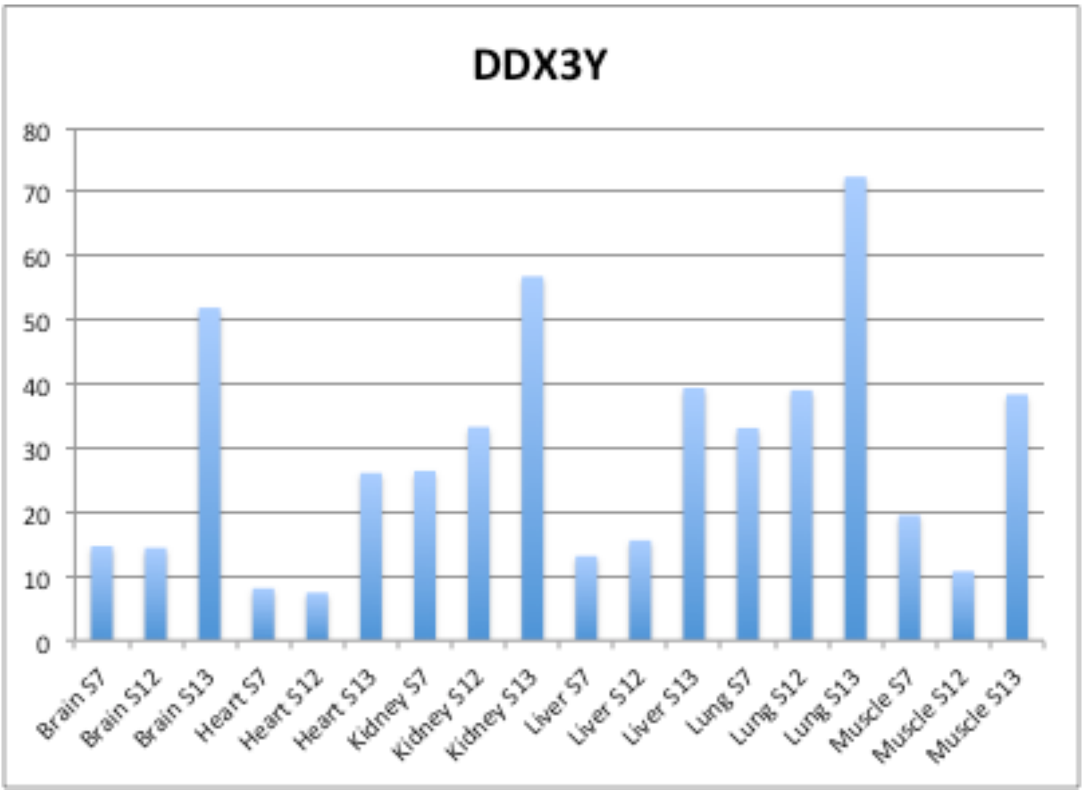


Figure 18. Expression levels of DDX3Y gene in the various samples, showing great variability in both individuals and tissues.

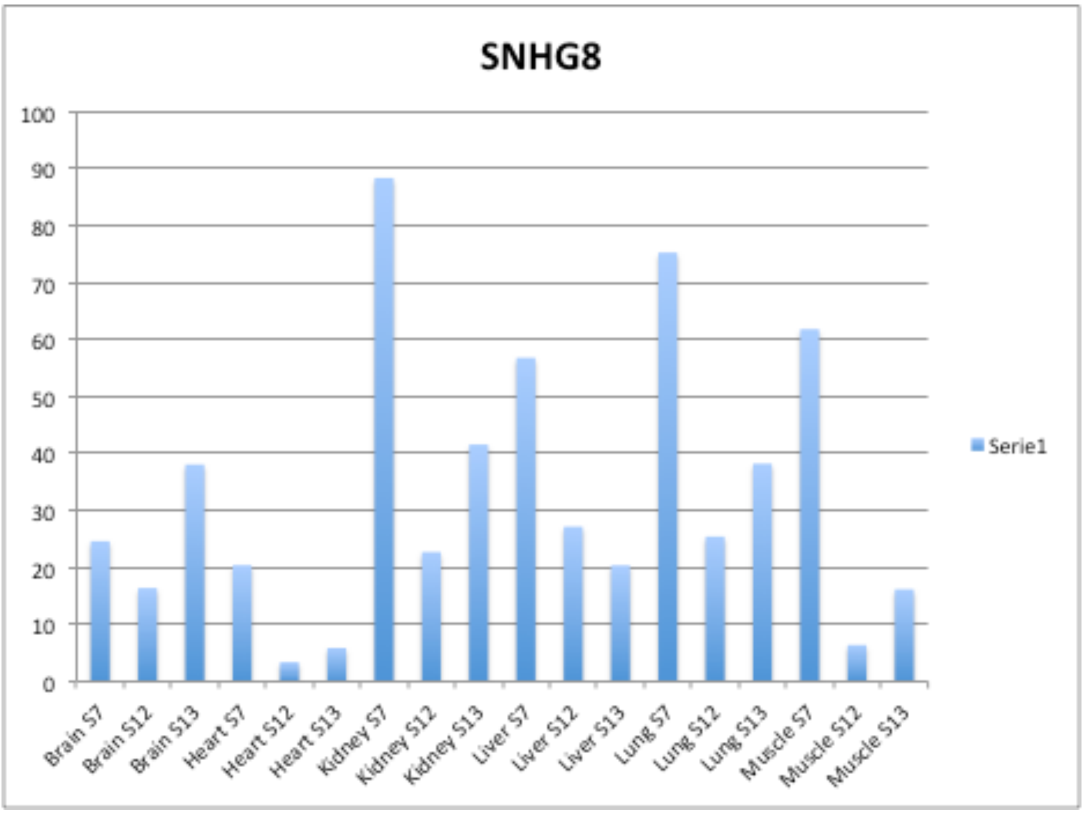
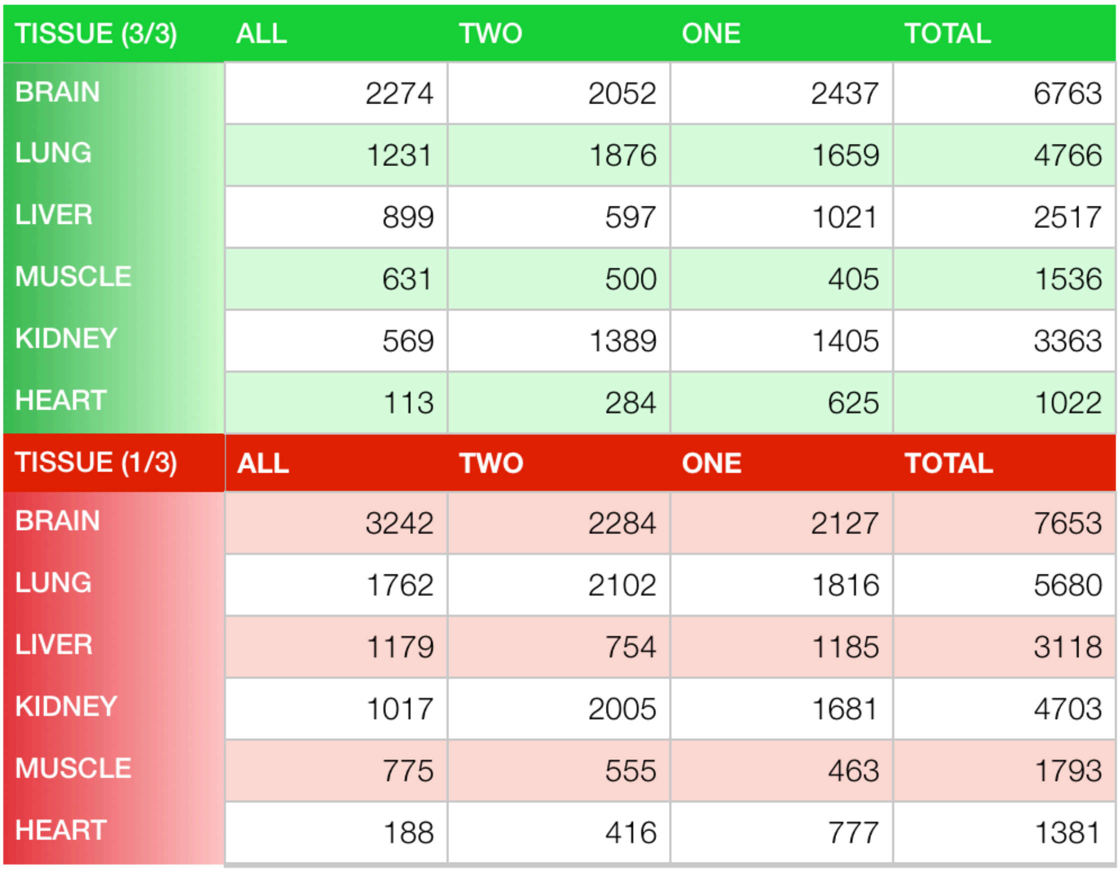


Figure 19. Expression levels of SNHG8 gene in the various samples, showing great variability in both individuals and tissues.

At the transcript level, by applying the general sample specificity test with a Benjamini-Hockberg corrected p-value threshold of 0.01, we obtained 15,931 transcripts of the 78,289 considered (20%) showed a significant deviation from the uniform distribution. If compared to the 68,557 with an expression value greater than zero in at least one sample the percentage rose to 23%, and to 35% if only transcripts with an expression FPKM value of at least one in at least one of the samples were considered (45,212 in all). At the gene level, results are consistent with the previous ones, with 11,118 genes out of 28,812 that had at least one transcript with expression significantly different from the uniform distribution across the samples studied.

Once again we deemed a transcript to be specific for a given individual in a given tissue if it passed the *single sample specificity test* in *all* the three replicates for the sample itself with a p-value threshold of 0.05, and to be tissue specific if specific for the tissue in at least one individual. Results shown in Table 7 quite obviously confirm the same trend observed at the gene level, with however an increase of the preponderance of transcripts specific for one or two individuals, confirming the complex expression patterns like the examples shown for the genes.



TISSUE (3/3)	ALL	TWO	ONE	TOTAL
BRAIN	2274	2052	2437	6763
LUNG	1231	1876	1659	4766
LIVER	899	597	1021	2517
MUSCLE	631	500	405	1536
KIDNEY	569	1389	1405	3363
HEART	113	284	625	1022

TISSUE (1/3)	ALL	TWO	ONE	TOTAL
BRAIN	3242	2284	2127	7653
LUNG	1762	2102	1816	5680
LIVER	1179	754	1185	3118
KIDNEY	1017	2005	1681	4703
MUSCLE	775	555	463	1793
HEART	188	416	777	1381

Table 7. Number of tissue specific transcripts in the six tissues detected in all three, two and one individual, in the case all three replicates (3/3, upper) or at least 1 replicate (1/3, lower) passed the specificity test.

The results of the test at the gene level and at the transcript level were obviously identical for all those genes with only one annotated transcript. Those with more than one alternative transcript could instead be split into different categories, summarized as follows:

a) All the annotated transcripts for the gene had the same sample-specificity of the gene itself.
b) Only some of the annotated transcripts for the gene had the same sample-specificity of the gene itself, while others did not present any significant bias.
c) The alternative transcripts combined together had the same specificity of the gene, but no transcript has exactly the same specificity of the gene.

d) A set of alternative transcripts had combined the same specificity of the gene, but further transcripts resulted to be specific for additional tissues.
e) The gene did not result sample-specific in any way according to the test, but some of its transcripts were.
f) The gene resulted to be sample-specific, but none of its transcripts, taken singularly, was.

Genes with all transcripts with the same tissue specificity (group a) were just a few hundreds, mainly because the majority had some of the alternative transcripts expressed at very low levels throughout all the samples investigated. This fact yielded the larger number of genes in group (b), in all comprising more than 6,000 out of the 9,083 significant genes with at least two alternative transcripts. Some of these genes nevertheless have one or more transcripts that are not sample specific, but with expression values high enough throughout the samples. The latter might thus be considered to be expressing both “housekeeping” and “tissue-specific” isoforms, as shown in the example of Figure 20.

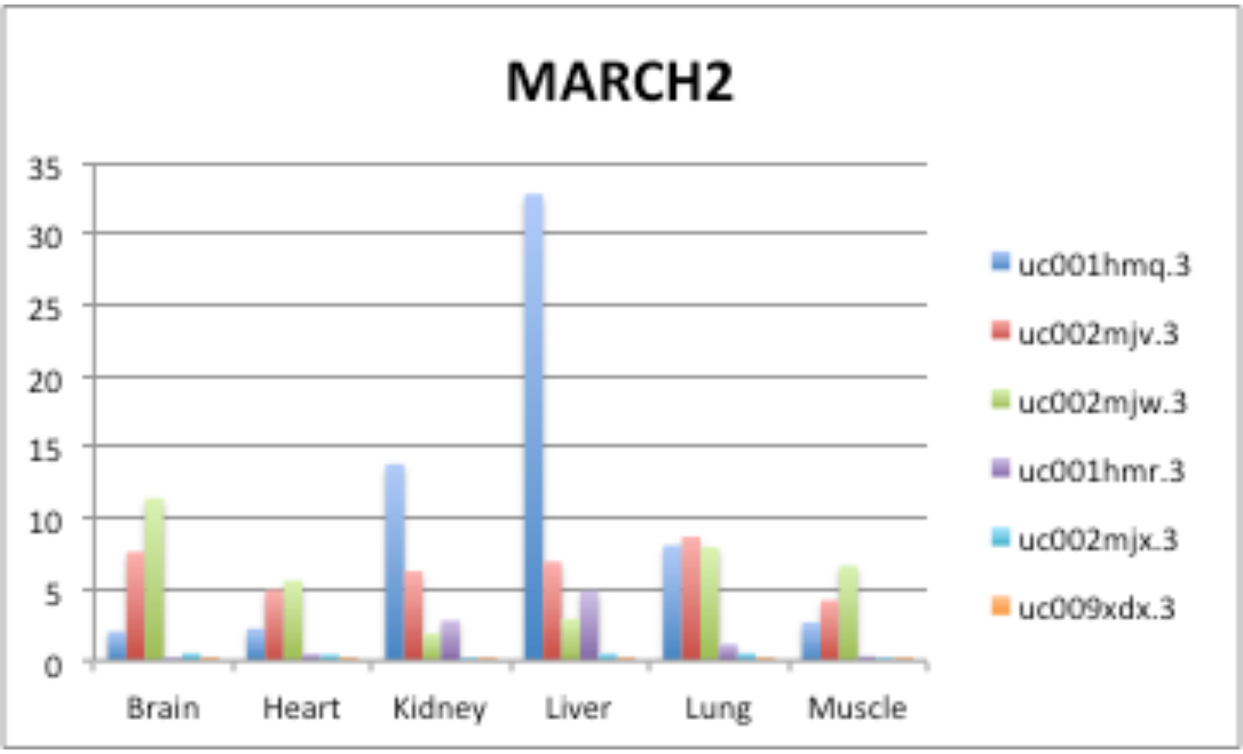


Figure 20. MARCH2 isoforms expression levels in the six tissues.

MARCH2 is specific for liver at the gene level, has a transcript specific for liver and kidney (blue), while the green and red isoforms do not pass the test and have an uniform expression level throughout the samples high enough to be considered to be housekeeping. Hence the gene can be considered to express both tissue specific and housekeeping isoforms (plot of the average transcript level in each tissue across all the replicates).

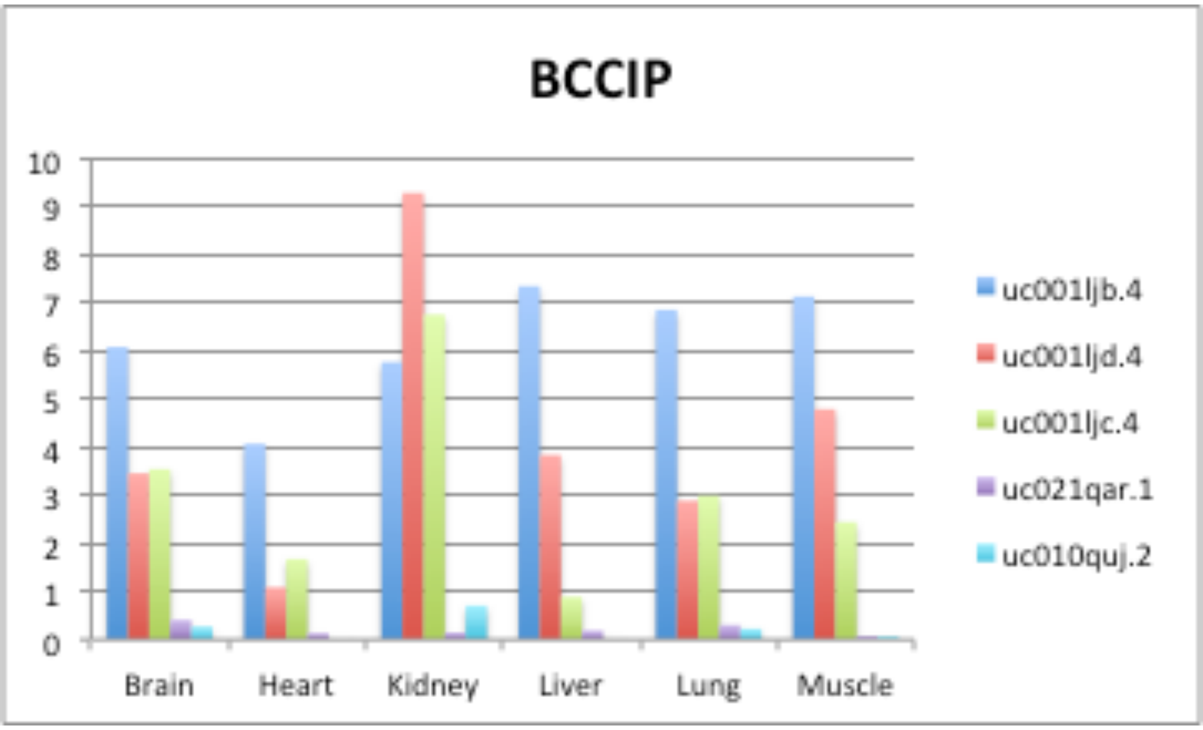


Figure 21. BCCIP isoforms expression levels in the six tissues.

BCCIP (Figure 21) has the main isoform (blue) that can be clearly considered housekeeping, with a minor isoform (red) over-expressed in kidney and kidney specific.

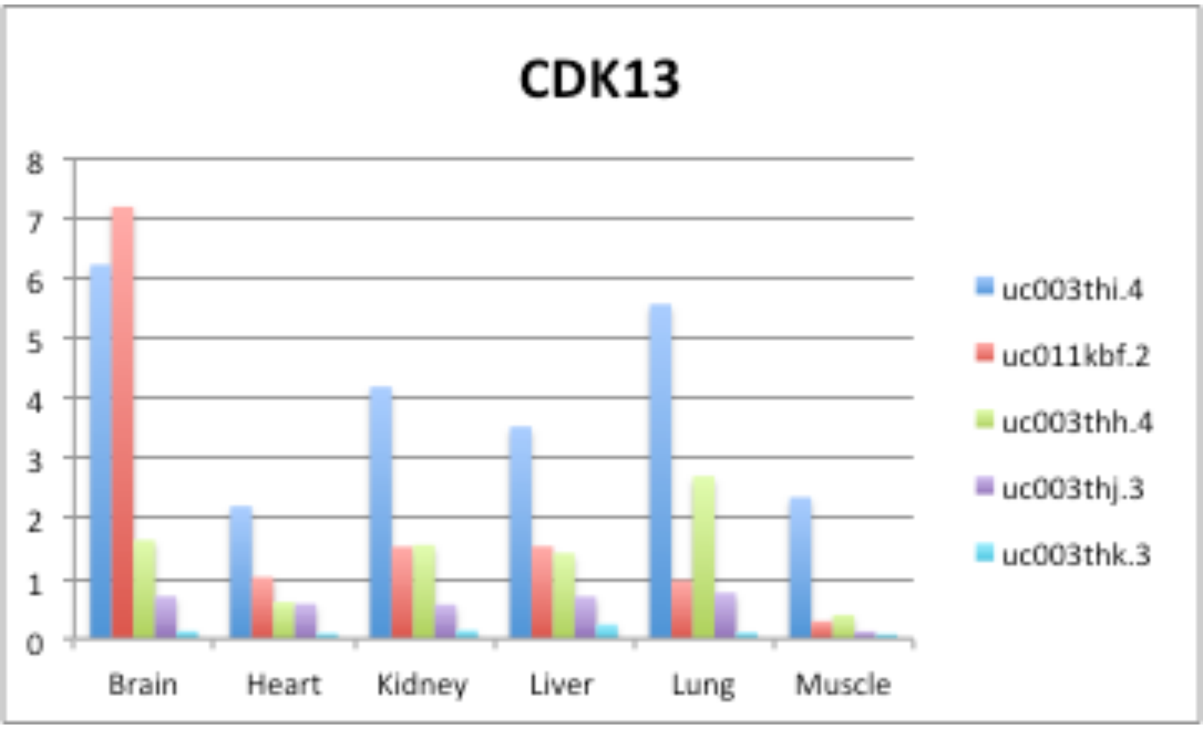


Figure 22. CDK13 isoforms expression levels in the six tissues.

CDK13 (Figure 22) has a brain specific transcript (red) and at least another one (blue) that does not pass the significance test and can be considered housekeeping.

While we can decide with our test whether a gene or transcript has an expression pattern significantly biased towards one or more samples, it is less straightforward to decide a criterion taking into account both expression values and their distribution across the samples to consider it “housekeeping”. To generalize the above examples we might define a transcript to be housekeeping with a quite stringent criterion: the p-value returned from the specificity test across the sample had to be greater than 0.1 (so to avoid “near misses”, that is, transcripts with biased expression patterns and p-values close to the threshold), and to be transcribed with FPKM greater than one in all the samples studied. In this way, a small, but remarkable group of about 200 genes then resulted to express both tissue specific and housekeeping isoforms, about 25% of which had the major isoform (the one with highest mean FPKM value) to be housekeeping.

Group (c) comprises about 1,500 genes which express different specific isoforms in different tissues, each one driving the specificity of the gene towards one or more of the samples.

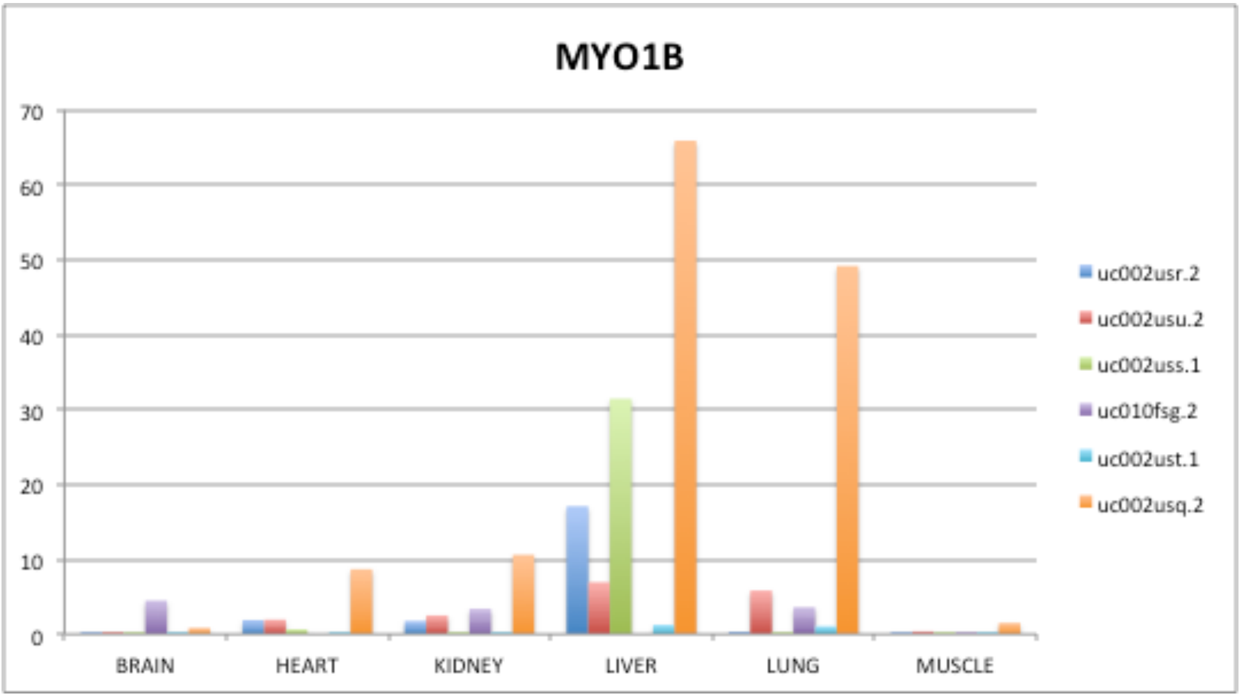


Figure 23. MYO1B isoforms expression levels in the six tissues.

Group (d) is comprised by a sizable number of 803 genes with one or more major tissue specific isoforms, that drive the overall tissue specificity of the gene, but that however have also minor isoforms specific for additional tissues, that cannot be identified by considering expression at the whole gene level. That is, the minor isoforms do not contribute enough to the overall expression of the gene to make it tissue-specific as a whole (as in Figure 23).

Group (e) is comprised of further 215 genes, that have an expression pattern similar to the genes with major housekeeping isoforms detected before that express also minor tissue-specific ones. But, in this case the group includes genes that show an overall uniform expression pattern across the samples, that however changes significantly the abundance of the relative transcripts, which taken singularly result to be tissue specific. This latter category of genes can be also nicely captured by applying the isoform switch test, that we will discuss later on.

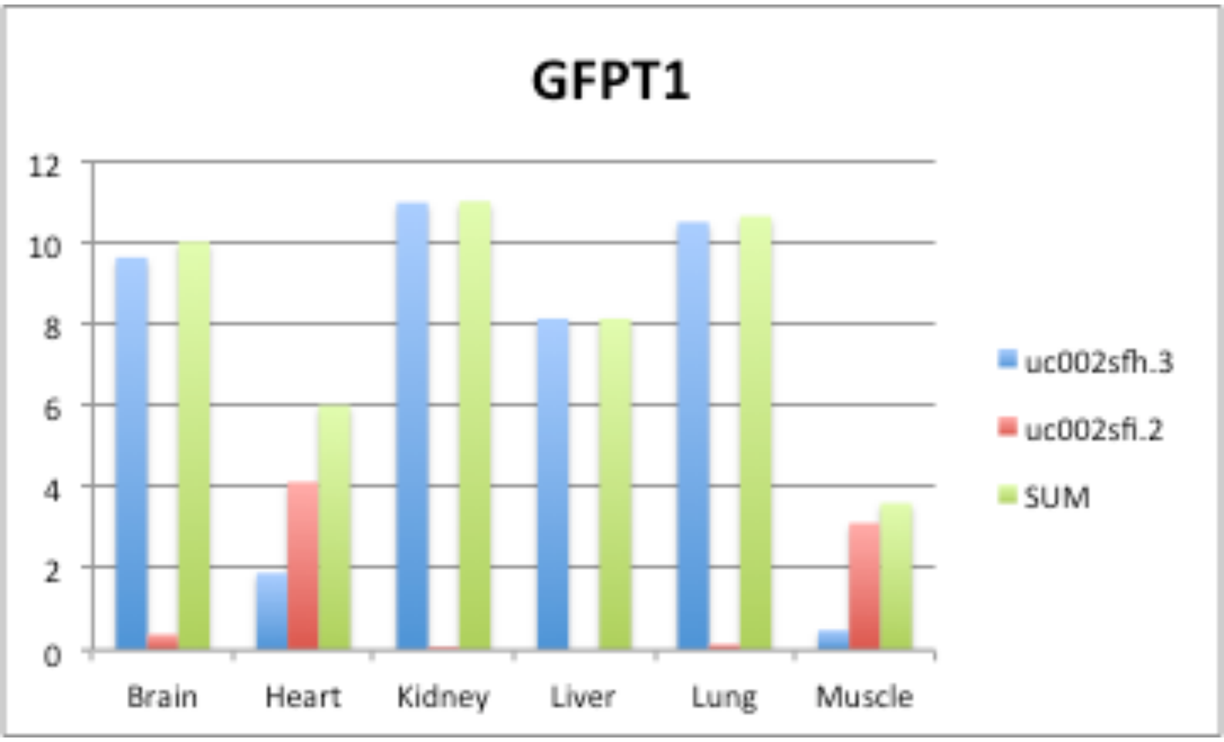


Figure 24. Comparision between GFPT1 isoforms levels and their expression values sum.

The overall expression of gene GFPT1 (green in Figure 24) does not change enough to make it specific for any tissue. The red transcript, however, is clearly heart- and muscle-specific. Notice also how the relative abundance of the blue and red transcripts switches significantly in these two tissues.

Finally, group (f) contains genes for which our results might seem to be contradictory, that is, tissue specific genes with no tissue specific transcripts. Rather, this can be seen as an additional feature of our method, that by considering simultaneously expression levels and bias in the distribution across the samples is able to identify single tissue-specific transcripts as seen before and, vice versa, genes for which the tissue specificity can be seen as a feature of the gene taken as a whole together with all its transcripts.

A non negligible total of 267 genes and 946 transcripts could not be considered to be specific for any sample after the single sample specificity test, because, although resulting to be significantly diverging from the uniform distribution in the general test, their variability within replicates resulted to be too high and did not pass the p-value threshold we set at the second step in all the three replicates in any of the samples. Unsurprisingly, among these we found the small/micro/tRNA genes, included in the UCSC annotation we employed, but for which the measures obtained from a total RNA-Seq experiment like the one we performed cannot be considered reliable, with a high variance of the estimated levels across the replicates. Hence, processing the data keeping replicates separate has also the side effect of identifying those genes or RNAs which should be eliminated from further downstream analyses because yielded inconsistent expression estimates.

All in all, one of the most striking results was the widespread presence of complex patterns of mixed individual- and/or tissue-biased gene and transcript expression. A large fraction of tissue specific genes and transcripts showed an expression pattern significantly biased towards samples coming from one or two individuals, and in general about one third of the transcripts that could be considered to be specific for one or more of the tissues studied could be considered to be so in only one individual. We obtained that 2007 genes and 4944 transcripts were to be considered "individual-specific", that is, with expression biased towards usually one, but in some cases two or more tissues coming from the same individual. In order to further investigate the extent of individual specific expression and its impact on our method, and at the same time on the usual assessments of tissue/sample specificity performed by averaging expression values across different samples, we reanalyzed the data in two more ways. First, we recomputed the general and sample specificity tests by considering the expression of a transcript in one tissue as the average of the expression values across the nine different samples and replicates of each tissue.

Results showed a lower number (9257) transcripts (11% of the total) passing the general sample specificity test with Benjamini-Hockberg corrected p-value threshold of 0.01 across the six tissues, that hence can be considered to be in general tissue-specific. The single sample specificity test revealed most of them (85%) to be specific for one tissue, 12% for two, and the rest (< 3%) for three or four tissues. As expected, all the transcripts that passed these two tests had been considered to significant also in the test performed keeping separate all the 54 samples and replicates. The lower number of transcripts detected to be significant by using FPKM averages was due to the fact that, as discussed in the previous section, when averaging over replicates we need in general transcripts to have higher expression values to be considered significantly tissue-specific. And, all variability among individuals is flattened by average values, that often were lowered and resulted to be non-significant. Thus, by processing all the samples and replicates separately from one another our method is able to highlight the details of cases in which genes and/or transcripts show a clear tissue specificity, but are at the same time highly variable across individuals.

To shed further light on the variability across individuals, we analyzed separately the nine samples coming from the same tissue, in order to detect once again whether there were significant variation across the individuals, but this time comparing only samples coming from the same tissue. We considered a transcript to be individual-specific if the test was passed in all the three replicates for the same individual. Results confirmed the widespread individual specific expression patterns detected in the first analysis performed across all the samples. We processed with the general sample specificity test the nine samples of each tissue with a Benjamini-Hockberg corrected p-value threshold of 0.01.

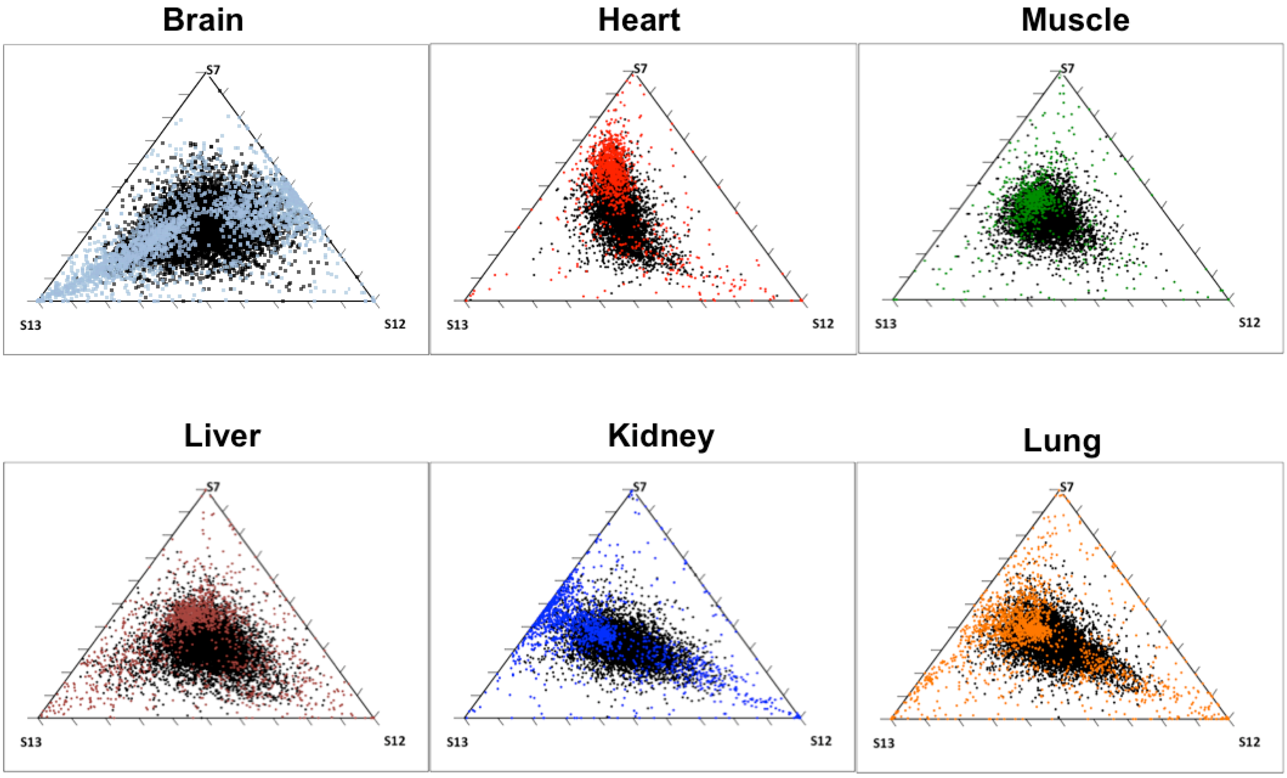
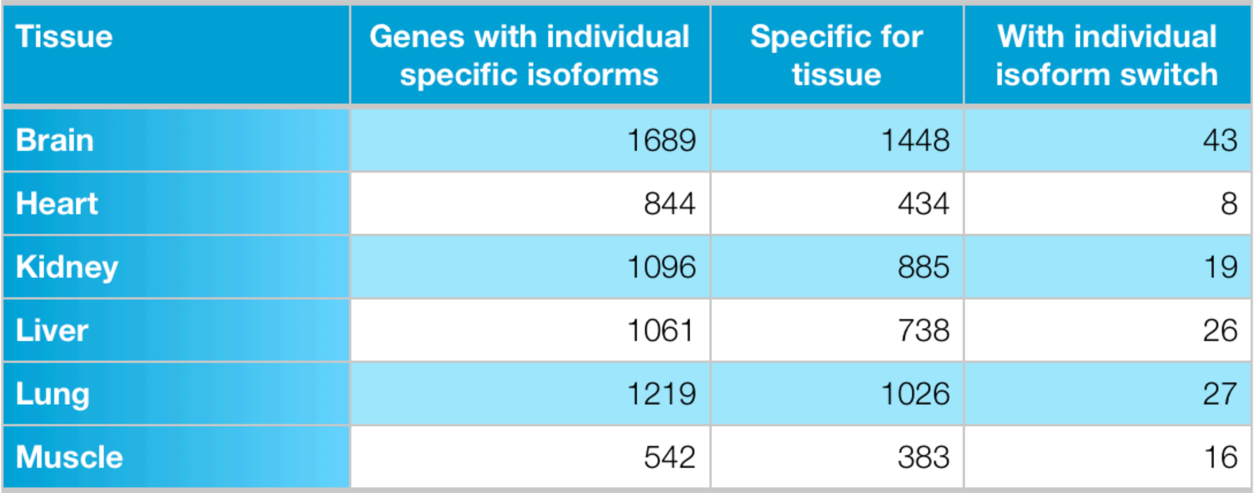


Figure 25. Triax plot showing the variability amng tissue samples, the color highlighted dots represent statistically significant differentially expressed genes.

The results are summarized in Figure 25, built by splitting the overall expression value of each transcript in each tissue into the contributions of each of the three individuals, and highlighting the transcripts yielding a significant variation in the samples in our test. It can be seen how the plot, in which each point corresponds to a transcript with FPKM > 1 in at least one individual, is clearly polarized in all the tissues towards the edges or the vertices of the triangle, instead of having all the points grouped in the middle as expected by samples with no variability. Overall, the test returned 2115 individual specific transcripts in brain, belonging to 1832 different genes, 994 (776 genes) in heart, 1365 (1096 genes) in kidney, 1337 (1060 genes) in liver, 1568 (1219 genes) in lung, and 658 (541 genes) in muscle. Thus, a significant fraction of the transcripts that were considered to be individual specific for the test across all the samples available remained so if the comparison was restricted to only those samples coming from the same tissue. In Figure 25, transcripts specific for one individual tend to fall in the plot next to one of the triangle vertices, but it can be clearly seen how a sizable number of transcripts, falling in between two vertices along the edge connecting them, are specific for two individuals out of three, or vice versa, is under-expressed in one individual. The polarity of the plot changes according to the tissues: in brain we can see a clear separation between S13 and S7/S12, that is, there seem to be transcripts both over- and under-expressed in S13 with respect to the other two individuals. Kidney and lung show a similar pattern, but in this case it is S12 differing from S7 and S13. Liver is similar to the latter two, but it is less polarized (fewer genes fall close to vertices or sides of the triangle) and seems to show a higher number of transcripts specific in each of the three individuals. Muscle has the least polarized expression pattern, with a lower number of transcripts with individual-specific expression towards S7 and S13. Finally, heart has a large number of transcripts over-expressed only in S7, in line with the observation that most of the heart-specific transcripts are so in only one individual.

Interestingly, not all the transcripts showing individual bias in a tissue belonged in turn to genes that were reported to have transcripts specific for the tissue itself. In other words, in brain we had 1689 genes with at least one transcript with individual biased expression, but only 1448 had transcripts considered brain specific in at least one individual by the general test (Table 8). The tissue with the highest number of individual biased genes not tissue specific was heart, where less than half of the 844 genes with individual specific transcripts was heart-specific.



Tissue	Genes with individual specific isoforms	Specific for tissue	With individual isoform switch
Brain	1689	1448	43
Heart	844	434	8
Kidney	1096	885	19
Liver	1061	738	26
Lung	1219	1026	27
Muscle	542	383	16

Table 8. Number of genes with individual-specific and tissue-specific isoforms, and with individual isoforms switches.

Another interesting feature that can be gathered from the table is that individual specific isoform switch (that is, a gene has to have at least two isoforms specific for different individuals) resulted to be a rare phenomenon, restricted to a few cases. That is, only a very limited subset of genes had alternative transcripts specific for different individuals, as shown for example for gene MPP6 in brain (brain specific) that shows a clear main isoform switch in the three individuals (Figure 26). MORF4L1, although not muscle specific, shows another clear switch between the main isoform expressed by S12, against that of S7 and S13 (Figure 27). Notice that in both cases the overall expression level of the gene is balanced among the individuals.

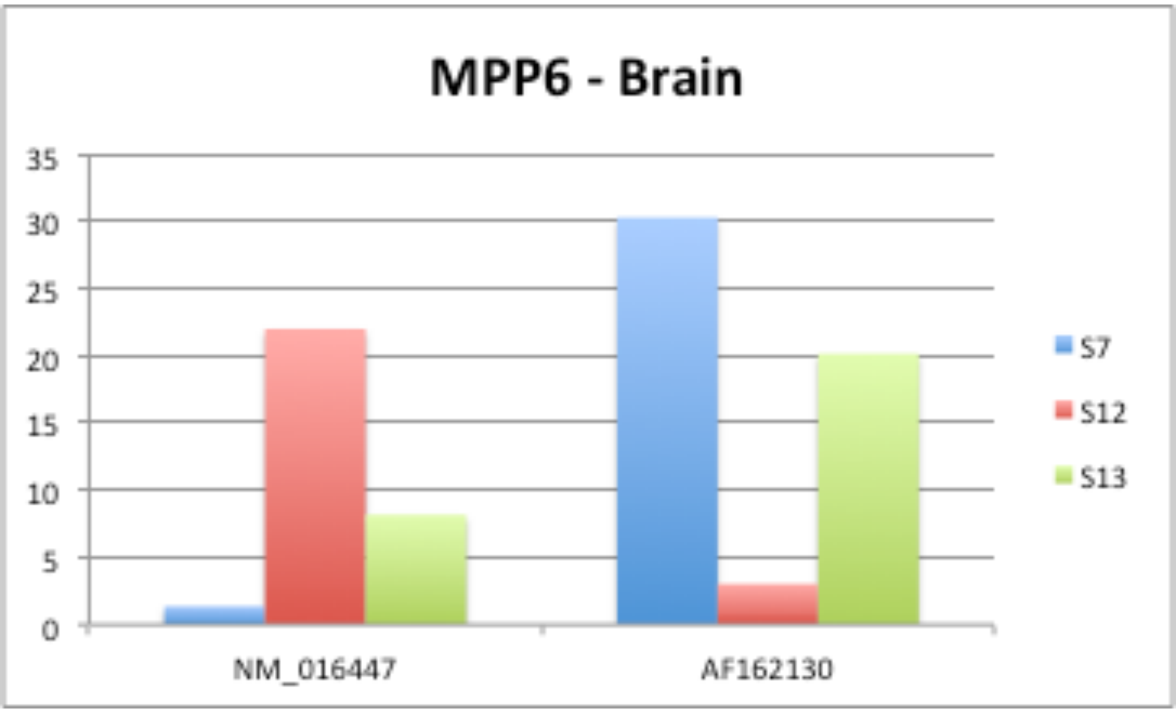


Figure 26. MPP6 expression levels in the three individuals (Brain).

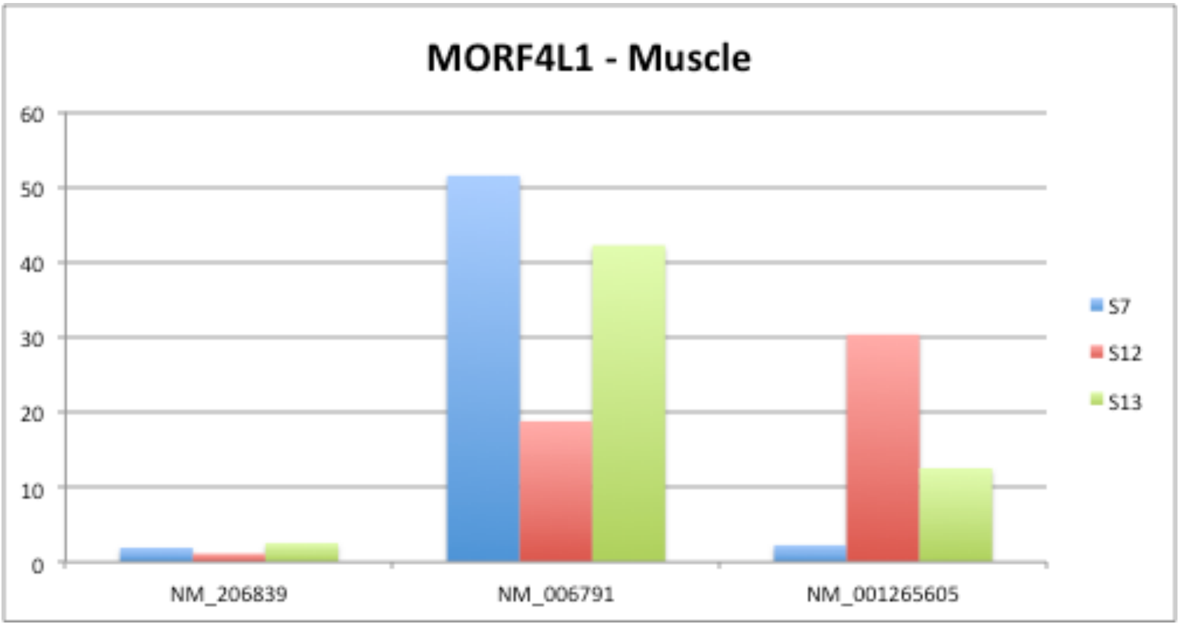


Figure 27. MORF4L1 expression levels in the three individuals (Muscle).

On the other hand, in the different tissues we noticed that genes with individual bias tended to belong to the same gene families, with individual specificity that seemed to be complementary. An example is the expression of synaptotagmin genes in brain in Figure 28. We can see in the plot, showing the split among the individuals of the overall expression of the genes, that while the cumulative expression of the members of the family is balanced among the three individuals (rightmost column), there are significant biases in most of the genes, that tend to be split between those specific for S7 and 12 (first 5-6 genes to the left) against those specific for 13 (last four genes to the right). In other words, the generation of “individual specific isoforms” is due to individual specific splicing to some extent, but more remarkably also with the individual specific expression of different paralogous genes of the same family (Figure 28).

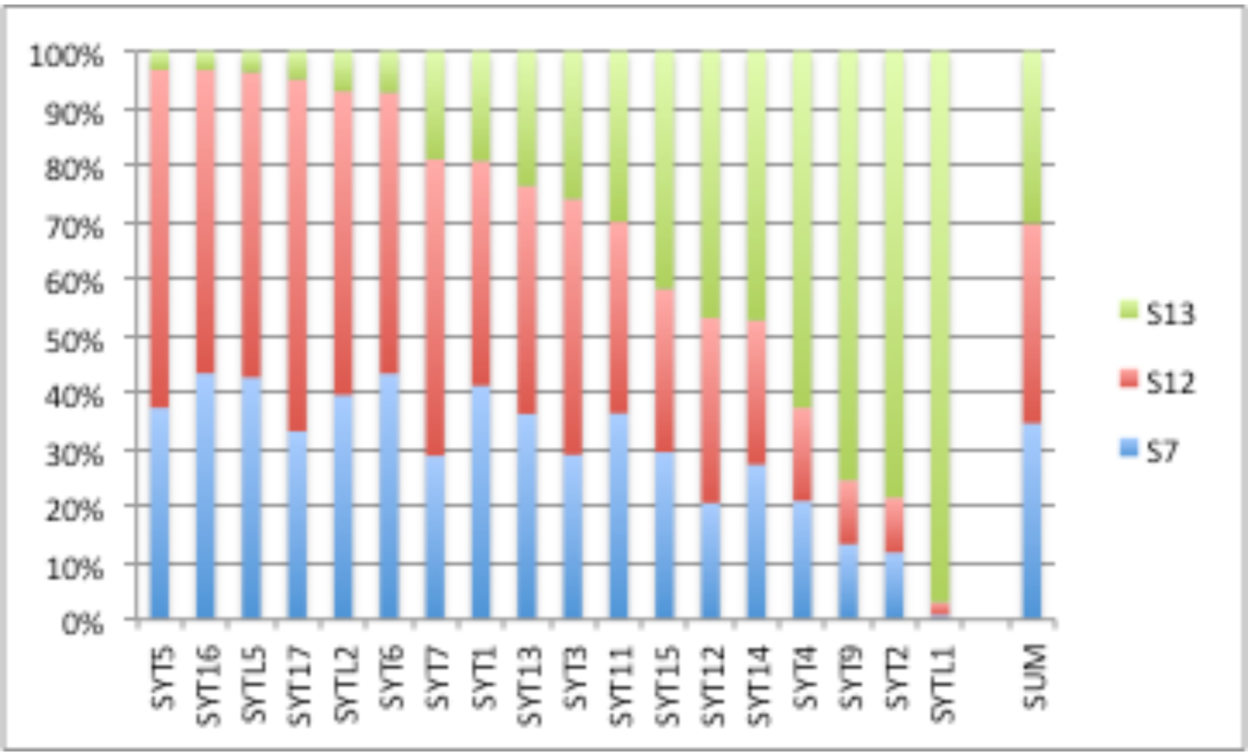


Figure 28. Expression of Synaptotagmin gene family in brain.

All in all, also this second analysis confirmed that transcripts that were detected to be as sample specific in the general analysis can be split into those that can be considered to be tissue-specific across the different individuals, those whose tissue specificity derives from individual specificity, and, finally, a significant subset of genes and transcripts resulting to have an individual specific bias in one or more tissues, but without any significant tissue specificity.

Then, we wanted to characterize genes and transcripts showing a tissue specific and/or individual specific expression patterns in the different tissues. For this analysis, we employed the functional annotation enrichment tools available at DAVID (Dennis et al., 2003), which includes and evaluates for a set of genes enrichment for Gene Ontology categories, annotated protein domains, pathways, and other annotations, performing a clustering of annotations that can be referred to similar functional categories.

Unsurprisingly, all the tissue-specific sets of genes returned by the initial sample specificity test across the 54 samples were enriched for functional annotations consistent with the respective tissues, and had their tissue specificity confirmed by the corresponding data available in DAVID (UP_TISSUE enrichment analysis).

A more interesting category was instead the set of genes with different splicing isoforms resulting to be specific for different tissues. Functional enrichment analysis returned “actin binding” and “cytoskeleton” to be the most relevant functional and localization annotations. This interesting finding supports the plausible correlation between cell/tissue type and cytoskeleton specificity mediated by alternative splicing differential regulation.

Switching to individual-specific gene sets, we started from brain, that showed the highest degree of individual-specific expression. Virtually the same categories, all related to neurons (synapse, axon, etc.), were found to be enriched in all the individuals. This is clearly the result of having genes of the same family with the same functional annotation differentially expressed in the individuals. Figure 29 shows the dispersion across the different individuals of the expression of transcripts belonging to genes with cellular component annotation “synapse” (light blue colored dots are transcripts passing the test, hence individual-specific). The plot clearly shows that the three individuals express different sub-classes of genes and isoforms associated with synapses, with the greater difference between S7/S12 and S13.

Another interesting case were genes whose transcripts over-expressed in individual S7, for which the functional analysis reported categories like ribosomal proteins, mitochondrion, and respiratory chain (Bonferroni corrected p-value < 10^{-50}), but also muscle contraction (p-value >10^{-15}), heart contraction (>10^{-7}), dilated cardiomyophathy (KEGG pathway enrichment >10^{-5}). Notice that most of the genes belonging to the first three categories did not result to be heart-specific.

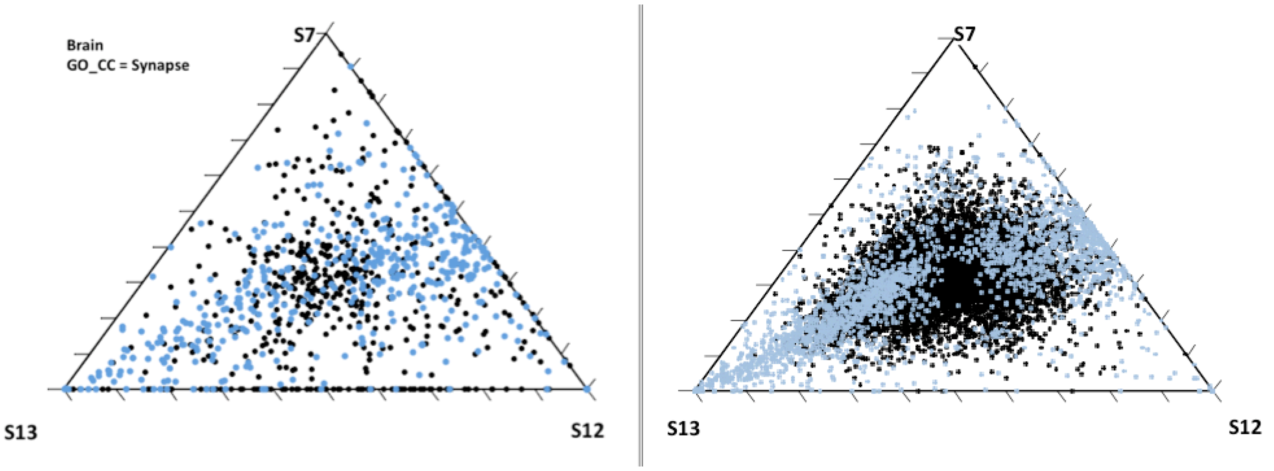


Figure 29. Dispersion across the different individuals of the expression of transcripts belonging to genes with cellular component annotation "synapse".

All in all, results point to the fact that in more than half of the cases tissue specific gene expression can be associated with significant variation throughout different individuals as well, while individual-specific alternative transcript and splicing usage was much less restricted. While in general tissue specific bias can be anyway recovered by pooling together all the samples, future transcriptome analyses should thus consider this very important aspect, and significantly different results can be obtained according to the source of the RNA samples analyzed and the conditions in which the sampling was performed.

Tissue-specific Splicing and Isoform Switches

We then better characterized genes with significant isoform switches across different samples. For sake of simplicity we present here the results obtained by pooling together all the samples and replicates for the same tissue and using in the analysis the resulting average FPKM value.

We applied the "isoform switch test" to the 15,877 genes comprising the UCSC gene annotation with at least two alternative transcripts. The result was that 2,328 genes switched significantly the relative abundance of their alternative transcripts in a sample-specific way according to our measure, by using a Benjamini-Hockberg corrected p-value threshold of 0.05. That is, produced alternative transcripts with significantly different tissue specificity. Remarkably 387 of these genes did not have any transcript resulting to be tissue-specific according to the general sample specificity test. Thus, the isoform switch test is able to capture finer cases, in which the “isoform switch” itself can be considered to be significant while the overall expression of the gene has little or no tissue specificity. As for tissue specific isoforms, also for this measure brain showed to be the sample with higher transcriptome variability, with 1211 genes with a significant isoform switch in brain compared to the other tissues, while the others had a comparable number (about 300-400) genes with a tissue specific switch.

For example, gene MAP3K4 (Figure 30) did not result to be tissue specific, but resulted to have a significant isoform switch in brain, with a change of major isoform.

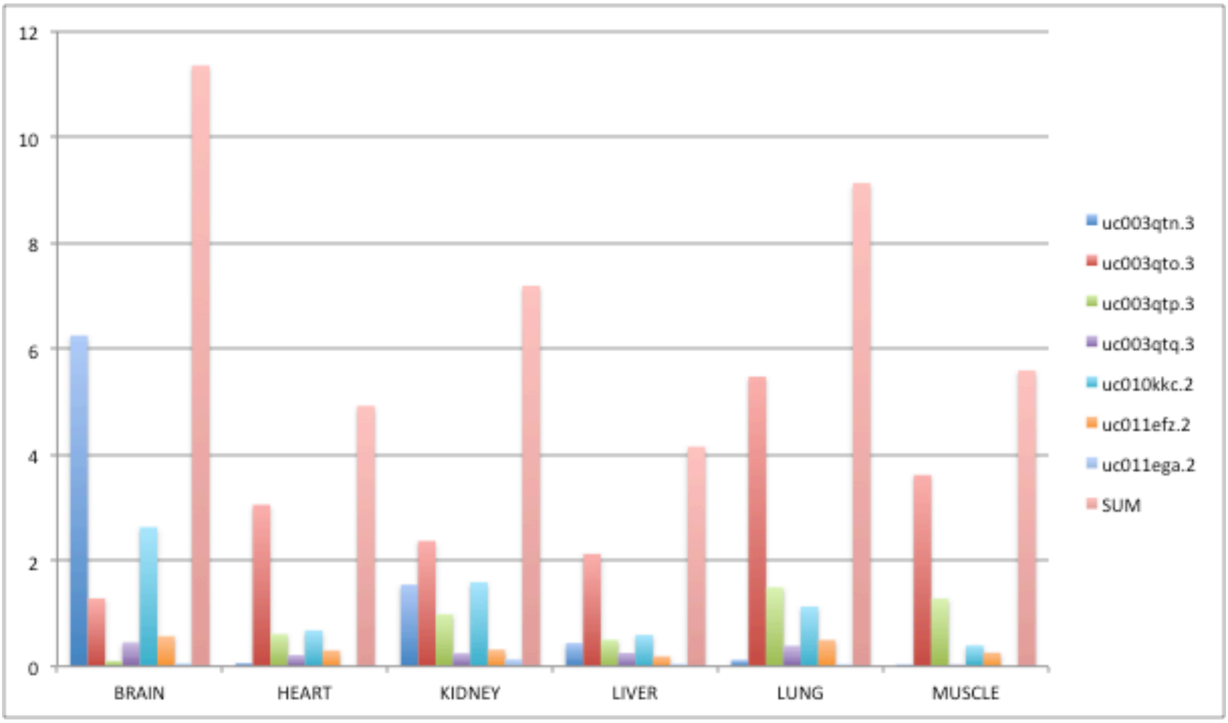


Figure 30. Expression of MAP3K4 isoforms across the six tissues.

Gene SH3PXD2A (Figure 31) is brain specific, where it expresses a different main isoform in brain (light blue) compared to the other tissues where the main isoform becomes the red one, while, the relative abundance of the purple and orange secondary isoforms does not show great difference across tissues.

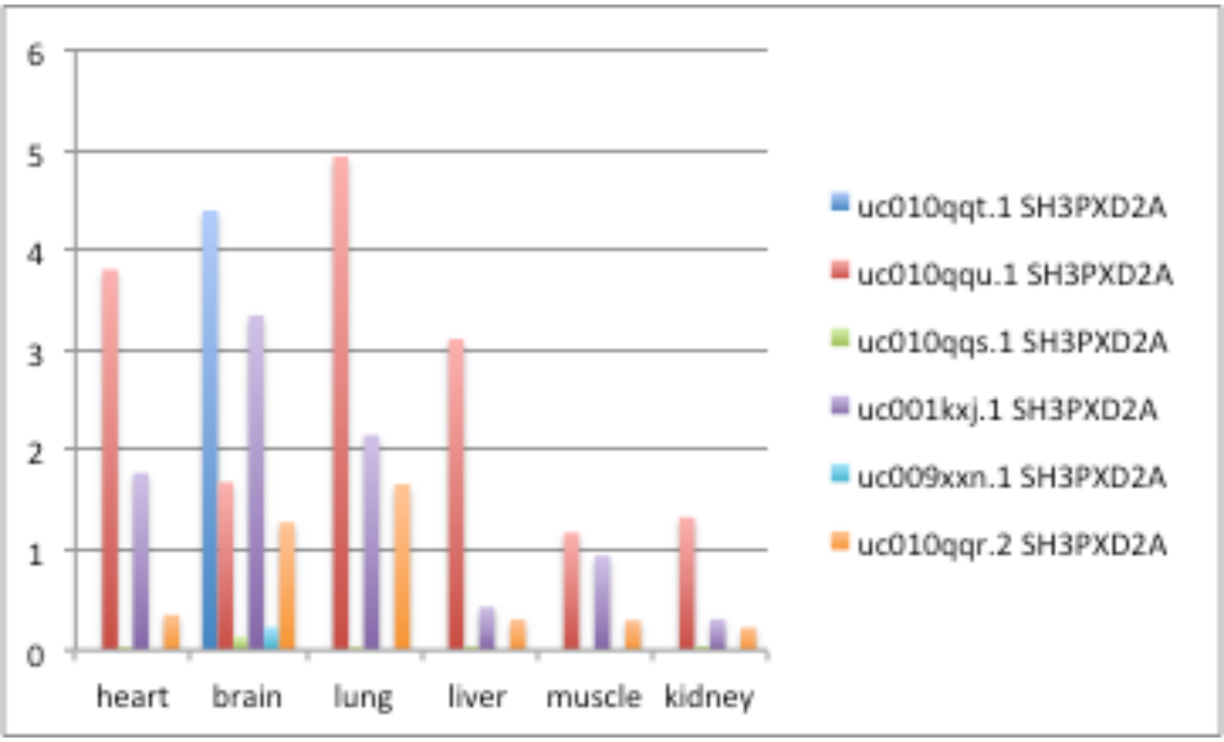


Figure 31. Expression of SH3PXD2A isoforms across the six tissues.

Gene DYM (Figure 32) is brain specific, mainly due to a secondary isoform expressed uniquely in brain.

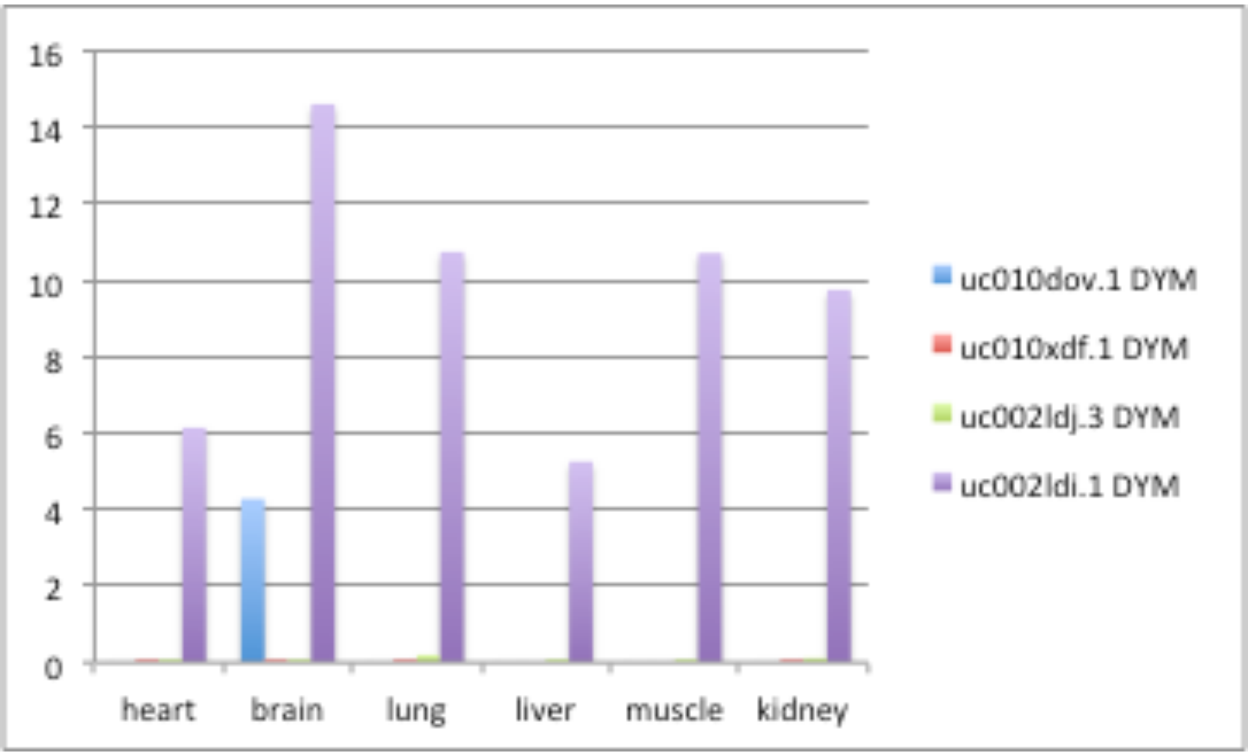


Figure 32. Expression of DYM isoforms across the six tissues.

Gene EML1 (Figure 33) is lung specific according to the general test, but has a completely different main isoform in brain, while the relative abundance of the others seems to remain unchanged.

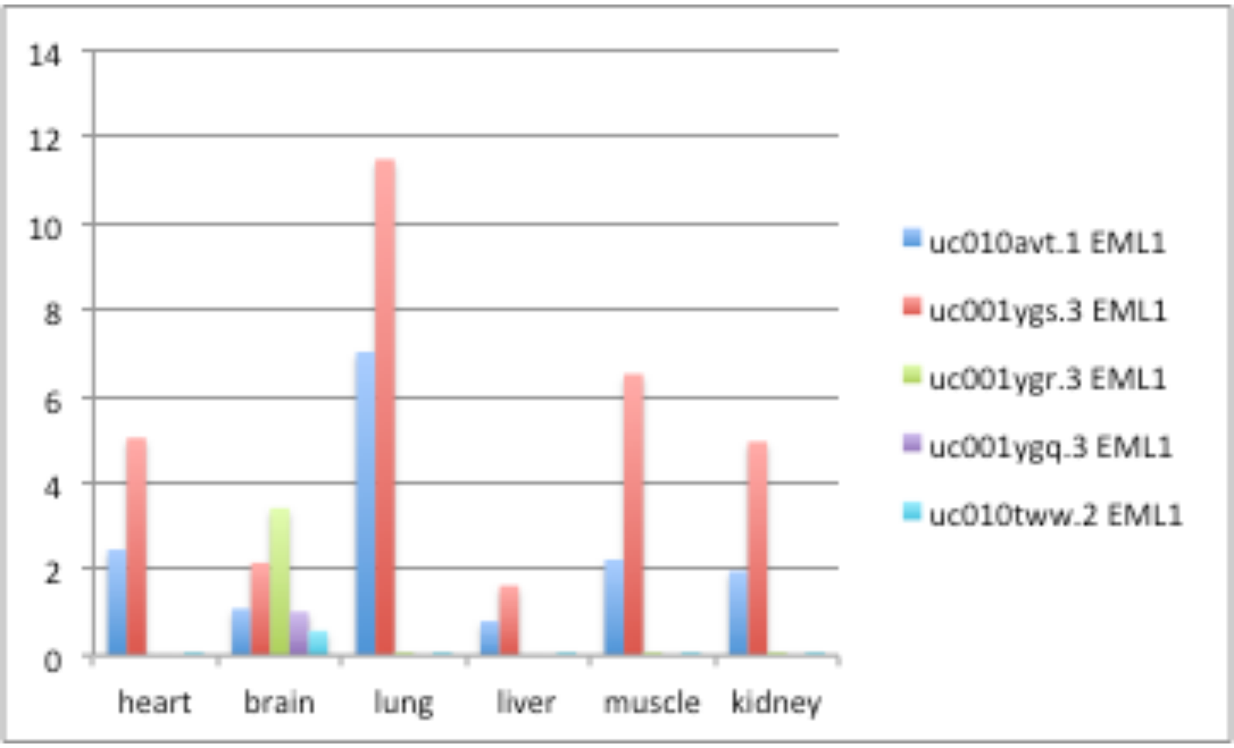


Figure 33. Expression of EML1 isoforms across the six tissues.

Gene WNK1 (Figure 34) presents complex changes of isoform abundance, changing relative abundance of three main isoforms across all the tissues investigated.

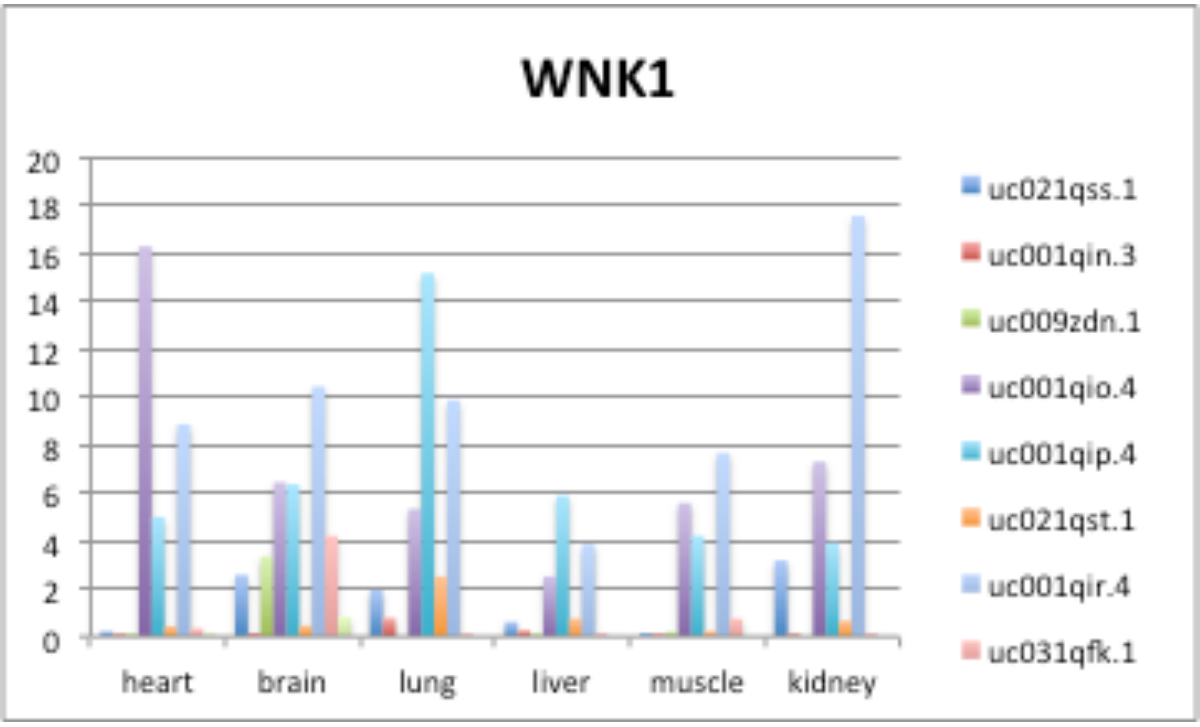


Figure 34. Expression of WNK1 isoforms across the six tissues.

Gene CLASP1 (Figure 35) shows three main isoforms, with relative abundance changing according to the tissue. Note that none of its transcripts alone passed the sample specificity test, while according to their overall expression the gene was reported as brain specific.

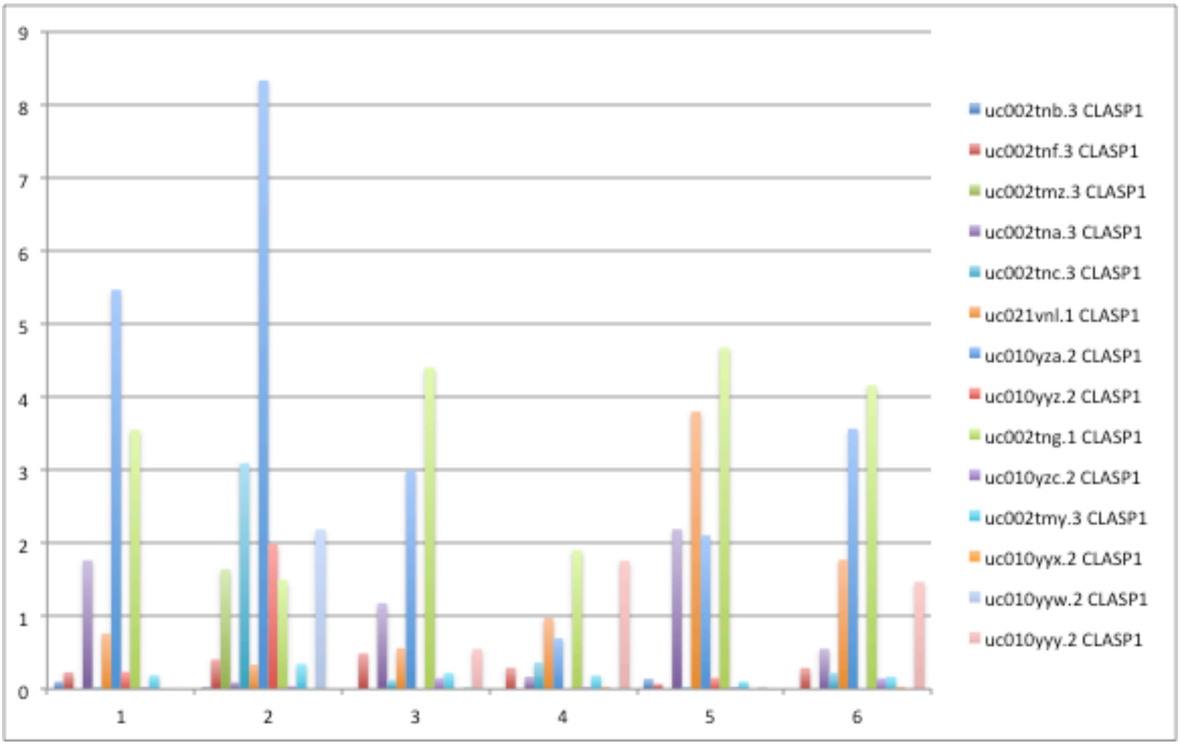


Figure 35. Expression of CLASP1 isoforms across the six tissues.

Finally, we computed the "exon specificity score" and the corresponding test on cassette exons, resulting by the comparison of the transcripts in the UCSC gene annotations and annotated as such in the UCSC genome browser track "Alternative Splicing Events". A total number of 4339 exons out of the 28351 that showed evidence of expression in at least one tissue resulted to be significant using a Benjamini-Hockberg corrected threshold of 0.05. Notice that the test does not evaluates tissue specificity per se, that is, not only if an exon belongs to a tissue-specific transcript, but also if there is a significant on/off switch of exon inclusion in different samples: in other words, a cassette exon is considered to be significant "switched" by our test if the same gene produces isoforms containing the exon with tissue specificity significantly different from those isoforms that do not include the exon. For example, in Figure 36, gene CLTA contains two cassette exons, which are clearly included only in brain, muscle, and marginally in kidney, yielding a significant switch in exon inclusion with respect to heart, lung, and liver. The topmost transcript, including both exons is brain-specific in all three individuals. At lower transcript levels, however, the inclusion of both exons is clearly specific also for kidney, with a significant switch with respect to the other tissues.

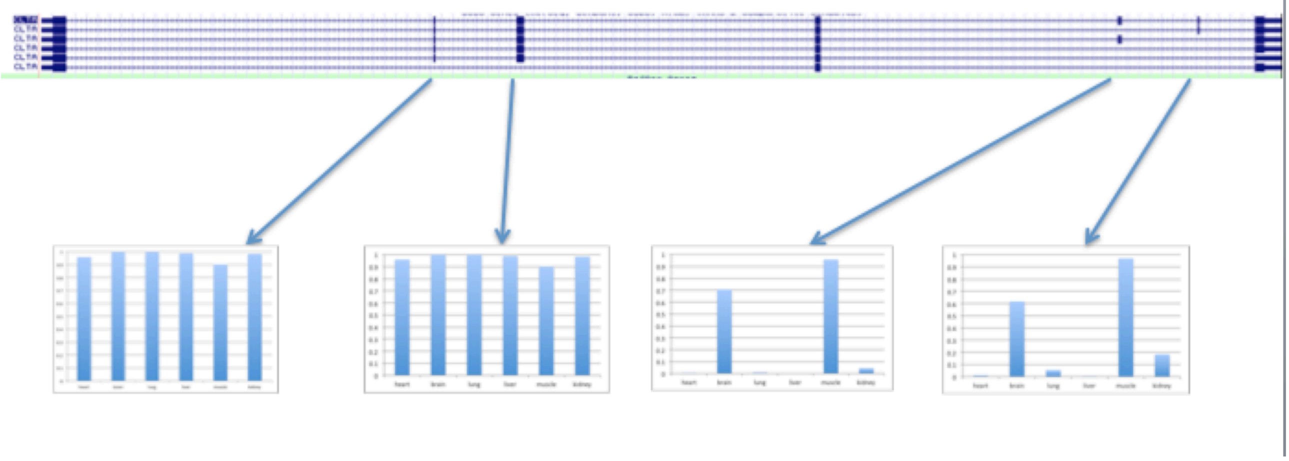


Figure 36. CLTA isoforms (from UCSC). The cassette-exon isoforms show a clear switch, being only expressed in brain, muscle and kidney, while the other isoforms show more or less the same levels in all the six tissues.

All in all, there are several examples showing how applying simultaneously all the different measures we introduced can provide a comprehensive view of the transcript/isoform complexity of a gene across different samples. In the examples we have shown they have been applied to the characterization of the variability of gene expression across different tissues and individuals. Other straightforward applications could be the analysis of tmle series experiments, where expression levels are estimated at different time points, e.g. developmental stages.

Concerning the data employed in this work, the next step will be the association of exome sequencing data to RNA-Seq data, in order to assess the impact of individual sequence variations to the individual specificity of gene expression, if any. And, we plan to take advantage of exome data in order to estimate the allele specificity of gene expression in the individuals. That is, heterozygous sequence variations identified by exome sequencing will be employed to assign transcripts to the corresponding alleles, to estimate the relative allele-specific abundance, and finally to assess whether the latter presents significant variation across the tissues and individuals for which data are available.

References

Aaij, C., & Borst, P. (1972). The gel electrophoresis of DNA. *Biochim Biophys Acta, 269*(2), 192-200.

Adams, J. (2008). Transcriptome: connecting the genome to gene function. *Nature Education, 1*(1), 195.

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J. J., Mayer, P., & Kawashima, E. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res, 28*(20), E87.

Agilent. (2015).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2

Ameur, A., Wetterbom, A., Feuk, L., & Gyllensten, U. (2010). Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol, 11*(3), R34. doi: 10.1186/gb-2010-11-3-r34

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol, 11*(10), R106. doi: 10.1186/gb-2010-11-10-r106

Andersson, S. G., Karlberg, O., Canback, B., & Kurland, C. G. (2003). On the origin of mitochondria: a genomics perspective. *Philos Trans R Soc Lond B Biol Sci, 358*(1429), 165-177; discussion 177-169. doi: 10.1098/rstb.2002.1193

Baggerly, K. A., Deng, L., Morris, J. S., & Aldaz, C. M. (2003). Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics, 19*(12), 1477-1483.

Baggerly, K. A., Deng, L., Morris, J. S., & Aldaz, C. M. (2004). Overdispersed logistic regression for SAGE: modelling multiple groups and covariates. *BMC Bioinformatics, 5*, 144. doi: 10.1186/1471-2105-5-144

Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., & Schnable, P. S. (2007). SNP discovery via 454 transcriptome sequencing. *Plant J, 51*(5), 910-918. doi: 10.1111/j.1365-313X.2007.03193.x

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., . . . Blencowe, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science, 338*(6114), 1587-1593. doi: 10.1126/science.1230612

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., . . . Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59. doi: 10.1038/nature07517

Berget, S. M., & Sharp, P. A. (1977). A spliced sequence at the 5'-terminus of adenovirus late mRNA. *Brookhaven Symp Biol*(29), 332-344.

Biosystems, A. (2015).

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem, 72*, 291-336. doi: 10.1146/annurev.biochem.72.121801.161720

Borst, P., & Grivell, L. A. (1978). The mitochondrial genome of yeast. *Cell, 15*(3), 705-723.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., . . . Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol, 18*(6), 630-634. doi: 10.1038/76469

Burden, C. J., Qureshi, S. E., & Wilson, S. R. (2014). Error estimates for the analysis of differential expression from RNA-seq count data. *PeerJ, 2*, e576. doi: 10.7717/peerj.576

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421. doi: 10.1186/1471-2105-10-421

Castellana, S., Romani, M., Valente, E. M., & Mazza, T. (2013). A solid quality-control analysis of AB SOLiD short-read sequencing data. *Brief Bioinform, 14*(6), 684-695. doi: 10.1093/bib/bbs048

Cavelier, L., Johannisson, A., & Gyllensten, U. (2000). Analysis of mtDNA copy number and composition of single mitochondrial particles using flow cytometry and PCR. *Exp Cell Res, 259*(1), 79-85. doi: 10.1006/excr.2000.4949

Central dogma reversed. (1970). *Nature, 226*(5252), 1198-1199.

Chen, Y. R., Zheng, Y., Liu, B., Zhong, S., Giovannoni, J., & Fei, Z. (2012). A cost-effective method for Illumina small RNA-Seq library preparation using T4 RNA ligase 1 adenylated adapters. *Plant Methods, 8*(1), 41. doi: 10.1186/1746-4811-8-41

Chow, L. C., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (2000). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. 1977. *Rev Med Virol, 10*(6), 362-371; discussion 355-366.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res, 38*(6), 1767-1771. doi: 10.1093/nar/gkp1137

Colgan, D. F., & Manley, J. L. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes Dev, 11*(21), 2755-2766.

Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol, 2010*, 853916. doi: 10.1155/2010/853916

Crick, F. H. (1958). On protein synthesis. *Symp Soc Exp Biol, 12*, 138-163.

Crow, M. T., Mani, K., Nam, Y. J., & Kitsis, R. N. (2004). The mitochondrial death pathway and cardiac myocyte apoptosis. *Circ Res, 95*(10), 957-970. doi: 10.1161/01.RES.0000148632.35500.d9

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., . . . Yu, J. (2010). A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics, 96*(5), 259-265. doi: 10.1016/j.ygeno.2010.07.010

D'Erchia, A. M., Atlante, A., Gadaleta, G., Pavesi, G., Chiara, M., De Virgilio, C., . . . Pesole, G. (2015). Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity. *Mitochondrion, 20*, 13-21. doi: 10.1016/j.mito.2014.10.005

Daly, A. K. (2010). Genome-wide association studies in pharmacogenomics. *Nat Rev Genet, 11*(4), 241-246. doi: 10.1038/nrg2751

Davila Lopez, M., & Samuelsson, T. (2008). Early evolution of histone mRNA 3' end processing. *RNA, 14*(1), 1-10. doi: 10.1261/rna.782308

De Bruijn, N. G. (1946). A Combinatorial Problem. *Koninklijke Nederlandse Akademie v. Wetenschappen 49: 758–764.*

de Leeuw, W. J., Slagboom, P. E., & Vijg, J. (1989). Quantitative comparison of mRNA levels in mammalian tissues: 28S ribosomal RNA level as an accurate internal control. *Nucleic Acids Res, 17*(23), 10137-10138.

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One, 8*(12), e85024. doi: 10.1371/journal.pone.0085024

Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol, 4*(5), P3.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., & Pagano, A. (2007). The expanding RNA polymerase III transcriptome. *Trends Genet, 23*(12), 614-622. doi: 10.1016/j.tig.2007.09.001

Draghici, S., Khatri, P., Eklund, A. C., & Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet, 22*(2), 101-109. doi: 10.1016/j.tig.2005.12.005

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., & Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A, 100*(15), 8817-8822. doi: 10.1073/pnas.1133470100

Edmonds, M., & Abrams, R. (1960). Polynucleotide biosynthesis: formation of a sequence of adenylate units from adenosine triphosphate by an enzyme from thymus nuclei. *J Biol Chem, 235*, 1142-1149.

Edwards, A., & Caskey, C. T. (1991). Closure strategies for random DNA sequencing. *Methods, 3*(1), 41-47. doi: http://dx.doi.org/10.1016/S1046-2023(05)80162-8

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends Genet, 29*(10), 569-574. doi: 10.1016/j.tig.2013.05.010

Emrich, S. J., Barbazuk, W. B., Li, L., & Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res, 17*(1), 69-73. doi: 10.1101/gr.5145806

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res, 8*(3), 186-194.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res, 8*(3), 175-185.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., & Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res, 34*(3), e22. doi: 10.1093/nar/gnj023

Fernandez-Silva, P., Enriquez, J. A., & Montoya, J. (2003). Replication and transcription of mammalian mitochondrial DNA. *Exp Physiol, 88*(1), 41-56.

Flaherty, B. L., Van Nieuwerburgh, F., Head, S. R., & Golden, J. W. (2011). Directional RNA deep sequencing sheds new light on the transcriptional response of Anabaena sp. strain PCC 7120 to combined-nitrogen deprivation. *BMC Genomics, 12*, 332. doi: 10.1186/1471-2164-12-332

Fullwood, M. J., Wei, C. L., Liu, E. T., & Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res, 19*(4), 521-532. doi: 10.1101/gr.074906.107

Furth, J. J., Hurwitz, J., & Anders, M. (1962). The role of deoxyribonucleic acid in ribonucleic acid synthesis. I. The purification and properties of ribonucleic acid polymerase. *J Biol Chem, 237*, 2611-2619.

Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T., & Martin, J. F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics, 12*, 245. doi: 10.1186/1471-2164-12-245

Gustafsson, A. B., & Gottlieb, R. A. (2008). Heart mitochondria: gates of life and death. *Cardiovasc Res, 77*(2), 334-343. doi: 10.1093/cvr/cvm005

Henze, K., & Martin, W. (2003). Evolutionary biology: essence of mitochondria. *Nature, 426*(6963), 127-128. doi: 10.1038/426127a

Hong, G. F. (1981). A method for sequencing single-stranded cloned DNA in both directions. *Biosci Rep, 1*(3), 243-252.

Housby, J. N., & Southern, E. M. (1998). Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Res, 26*(18), 4259-4266.

Illumina. (2015). from http://www.illumina.com/

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., . . . Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science, 302*(5653), 2141-2144. doi: 10.1126/science.1090100

Kalia, A., & Gupta, R. P. (2005). Proteomics: A Paradigm Shift. *Critical Reviews in Biotechnology, 25*(4), 173-198. doi: doi:10.1080/07388550500365102

Kornberg, R. D. (1999). Eukaryotic transcriptional control. *Trends Cell Biol, 9*(12), M46-49.

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res, 15*(20), 8125-8148.

Kullback, S. L., R. A. (1951). On information and sufficiency. *Ann. Math. Statistics 22. 79–86.*

Lehman, I. R., Bessman, M. J., Simms, E. S., & Kornberg, A. (1958). Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli. *J Biol Chem, 233*(1), 163-170.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi: 10.1186/1471-2105-12-323

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform, 11*(5), 473-483. doi: 10.1093/bib/bbq015

Li, X., Nair, A., Wang, S., & Wang, L. (2015). Quality control of RNA-seq experiments. *Methods Mol Biol, 1269*, 137-146. doi: 10.1007/978-1-4939-2291-8_8

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J., & Deng, H. W. (2011). Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics, 27*(15), 2031-2037. doi: 10.1093/bioinformatics/btr319

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., & Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nat Genet, 21*(1 Suppl), 20-24. doi: 10.1038/4447

Lu, J., Tomfohr, J. K., & Kepler, T. B. (2005). Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics, 6*, 165. doi: 10.1186/1471-2105-6-165

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet, 24*(3), 133-141. doi: 10.1016/j.tig.2007.12.007

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380. doi: 10.1038/nature03959

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nat Rev Genet, 12*(10), 671-682. doi: 10.1038/nrg3068

Marziali, A., & Akeson, M. (2001). New DNA sequencing methods. *Annu Rev Biomed Eng, 3*, 195-223. doi: 10.1146/annurev.bioeng.3.1.195

Matlin, A. J., Clark, F., & Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol, 6*(5), 386-398. doi: 10.1038/nrm1645

Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., . . . Terauchi, R. (2005). SuperSAGE. *Cell Microbiol, 7*(1), 11-18. doi: 10.1111/j.1462-5822.2004.00478.x

Matsumura, H., Yoshida, K., Luo, S., Kimura, E., Fujibe, T., Albertyn, Z., . . . Terauchi, R. (2010). High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS One, 5*(8), e12010. doi: 10.1371/journal.pone.0012010

Maxam, A. M., & Gilbert, W. (1992). A new method for sequencing DNA. 1977. *Biotechnology, 24*, 99-103.

Mead, M. N. (2007). Nutrigenomics: the genome--food interface. *Environ Health Perspect, 115*(12), A582-589.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods, 5*(7), 621-628. doi: 10.1038/nmeth.1226

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol, 51 Pt 1*, 263-273.

Myers, T. W., & Gelfand, D. H. (1991). Reverse transcription and DNA amplification by a Thermus thermophilus DNA polymerase. *Biochemistry, 30*(31), 7661-7666.

Ng, P., Wei, C. L., & Ruan, Y. (2007). Paired-end diTagging for transcriptome and genome analysis. *Curr Protoc Mol Biol, Chapter 21*, Unit 21 12. doi: 10.1002/0471142727.mb2112s79

NHGRI. (2015). from http://www.genome.gov/

Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., & Barron, A. E. (2011). Landscape of next-generation sequencing technologies. *Anal Chem, 83*(12), 4327-4341. doi: 10.1021/ac2010857

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet, 12*(6), 443-451. doi: 10.1038/nrg2986

Nyren, P. (2007). The history of pyrosequencing. *Methods Mol Biol, 373*, 1-14. doi: 10.1385/1-59745-377-3:1

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet, 40*(12), 1413-1415. doi: 10.1038/ng.259

Pesole, G., Allen, J. F., Lane, N., Martin, W., Rand, D. M., Schatz, G., & Saccone, C. (2012). The neglected genome. *EMBO Rep, 13*(6), 473-474. doi: 10.1038/embor.2012.57

Picardi, E., & Pesole, G. (2012). Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat Methods, 9*(6), 523-524. doi: 10.1038/nmeth.2029

Poptsova, M. S., Il'icheva, I. A., Nechipurenko, D. Y., Panchenko, L. A., Khodikov, M. V., Oparina, N. Y., . . . Grokhovsky, S. L. (2014). Non-random DNA fragmentation in next-generation sequencing. *Sci Rep, 4*, 4532. doi: 10.1038/srep04532

Quail, M. A. (2001). DNA: Mechanical Breakage *eLS*: John Wiley & Sons, Ltd.

Quiagen. (2015a). from http://www.qiagen.com/it/products/catalog/sample-technologies/dna-sample-technologies/genomic-dna/dneasy-blood-and-tissue-kit/

Quiagen. (2015b).

Rackham, O., Shearwood, A. M., Mercer, T. R., Davies, S. M., Mattick, J. S., & Filipovska, A. (2011). Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. *RNA, 17*(12), 2085-2093. doi: 10.1261/rna.029405.111

Richterich, P. (1998). Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res, 8*(3), 251-259.

Roach, J. C., Boysen, C., Wang, K., & Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics, 26*(2), 345-353.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol, 12*(3), R22. doi: 10.1186/gb-2011-12-3-r22

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., . . . Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nat Methods, 7*(11), 909-912. doi: 10.1038/nmeth.1517

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi: 10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol, 11*(3), R25. doi: 10.1186/gb-2010-11-3-r25

Roche. (2015).

Rogers, Y. H., & Venter, J. C. (2005). Genomics: massively parallel sequencing. *Nature, 437*(7057), 326-327. doi: 10.1038/437326a

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., & Nyren, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem, 242*(1), 84-89. doi: 10.1006/abio.1996.0432

Ronaghi, M., Uhlen, M., & Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science, 281*(5375), 363, 365.

Rorbach, J., & Minczuk, M. (2012). The post-transcriptional life of mammalian mitochondrial RNA. *Biochem J, 444*(3), 357-373. doi: 10.1042/BJ20112208

Russell, J., & Zomerdijk, J. C. (2006). The RNA polymerase I transcription machinery. *Biochem Soc Symp*(73), 203-216.

Sambrook, J., & Russell, D. W. (2006a). Fragmentation of DNA by nebulization. *CSH Protoc, 2006*(4). doi: 10.1101/pdb.prot4539

Sambrook, J., & Russell, D. W. (2006b). Fragmentation of DNA by sonication. *CSH Protoc, 2006*(4). doi: 10.1101/pdb.prot4538

Sammeth, M., Foissac, S., & Guigo, R. (2008). A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol, 4*(8), e1000147. doi: 10.1371/journal.pcbi.1000147

Sanger, F., Nicklen, S., & Coulson, A. R. (1992). DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology, 24*, 104-108.

Sawicki, M. P., Samara, G., Hurwitz, M., & Passaro Jr, E. (1993). Human Genome Project. *The American Journal of Surgery, 165*(2), 258-264. doi: http://dx.doi.org/10.1016/S0002-9610(05)80522-7

Scarpulla, R. C. (1997). Nuclear control of respiratory chain expression in mammalian cells. *J Bioenerg Biomembr, 29*(2), 109-119.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Hum Mol Genet, 19*(R2), R227-240. doi: 10.1093/hmg/ddq416

Schneider, M. V., & Orchard, S. (2011). Omics technologies, data and bioinformatics principles. *Methods Mol Biol, 719*, 3-30. doi: 10.1007/978-1-61779-027-0_1

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., . . . Ragg, T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol, 7*, 3. doi: 10.1186/1471-2199-7-3

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics, 28*(8), 1086-1092. doi: 10.1093/bioinformatics/bts094

Shalon, D., Smith, S. J., & Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res, 6*(7), 639-645.

Shannon, C. E., Weaver, W. (1949). The Mathematical Theory of Communication. *Univ of Illinois Press*.

Shendure, J. (2008). The beginning of the end for microarrays? *Nat Methods, 5*(7), 585-587. doi: 10.1038/nmeth0708-585

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol, 26*(10), 1135-1145. doi: 10.1038/nbt1486

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., . . . Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science, 309*(5741), 1728-1732. doi: 10.1126/science.1117389

Shine, J., & Dalgarno, L. (1975). Determinant of cistron specificity in bacterial ribosomes. *Nature, 254*(5495), 34-38.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., . . . Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A, 100*(26), 15776-15781. doi: 10.1073/pnas.2136655100

Smith, A. D., Xuan, Z., & Zhang, M. Q. (2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics, 9*, 128. doi: 10.1186/1471-2105-9-128

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol, 147*(1), 195-197.

Stangegaard, M., Dufva, I. H., & Dufva, M. (2006). Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. *Biotechniques, 40*(5), 649-657.

Stein, L. D. (2011). An introduction to the informatics of "next-generation" sequencing. *Curr Protoc Bioinformatics, Chapter 11*, Unit 11 11. doi: 10.1002/0471250953.bi1101s36

Sumner, A. T., de la Torre, J., & Stuppia, L. (1993). The distribution of genes on chromosomes: a cytological approach. *J Mol Evol, 37*(2), 117-122.

Taanman, J. W. (1999). The mitochondrial genome: structure, transcription, translation and replication. *Biochim Biophys Acta, 1410*(2), 103-123.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., . . . Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods, 6*(5), 377-382. doi: 10.1038/nmeth.1315

ThermoFisher. (2015). from http://www.thermoscientific.com/content/tfs/en/product/nanodrop-2000-2000c-spectrophotometers.html

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics, 25*(9), 1105-1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol, 28*(5), 511-515. doi: 10.1038/nbt.1621

Turcatti, G., Romieu, A., Fedurco, M., & Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res, 36*(4), e25. doi: 10.1093/nar/gkn021

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science, 270*(5235), 484-487.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351. doi: 10.1126/science.1058040

Wall, J. D., Tang, L. F., Zerbe, B., Kvale, M. N., Kwok, P. Y., Schaefer, C., & Risch, N. (2014). Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res, 24*(11), 1734-1739. doi: 10.1101/gr.168393.113

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63. doi: 10.1038/nrg2484

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. J.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953. *Nature, 248*(5451), 765.

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics, 26*(7), 873-881. doi: 10.1093/bioinformatics/btq057

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., . . . Wang, J. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics, 30*(12), 1660-1666. doi: 10.1093/bioinformatics/btu077

Yakovchuk, P., Protozanova, E., & Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res, 34*(2), 564-574. doi: 10.1093/nar/gkj454

Yassour, M., Kaplan, T., Fraser, H. B., Levin, J. Z., Pfiffner, J., Adiconis, X., . . . Regev, A. (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci U S A, 106*(9), 3264-3269. doi: 10.1073/pnas.0812841106

Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *J Genet Genomics, 38*(3), 95-109. doi: 10.1016/j.jgg.2011.02.003

Papers and Manuscripts

W510–W515 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 4 June 2012*
doi:10.1093/nar/gks483

Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets

Federico Zambelli[1], Gian Marco Prazzoli[1], Graziano Pesole[2,3] and Giulio Pavesi[1,*]

[1]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Italy, [2]Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy and [3]Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università di Bari, Italy

Received February 13, 2012; Revised April 25, 2012; Accepted May 2, 2012

ABSTRACT

The regulation of transcription of eukaryotic genes is a very complex process, which involves interactions between transcription factors (TFs) and DNA, as well as other epigenetic factors like histone modifications, DNA methylation, and so on, which nowadays can be studied and characterized with techniques like ChIP-Seq. Cscan is a web resource that includes a large collection of genome-wide ChIP-Seq experiments performed on TFs, histone modifications, RNA polymerases and others. Enriched peak regions from the ChIP-Seq experiments are crossed with the genomic coordinates of a set of input genes, to identify which of the experiments present a statistically significant number of peaks within the input genes' loci. The input can be a cluster of co-expressed genes, or any other set of genes sharing a common regulatory profile. Users can thus single out which TFs are likely to be common regulators of the genes, and their respective correlations. Also, by examining results on promoter activation, transcription, histone modifications, polymerase binding and so on, users can investigate the effect of the TFs (activation or repression of transcription) as well as of the cell or tissue specificity of the genes' regulation and expression. The web interface is free for use, and there is no login requirement. Available at: http://www.beaconlab.it/cscan.

INTRODUCTION

The regulation of eukaryotic gene transcription is a very complex process, which depends on interactions between transcription factors (TFs) and DNA, as well as on chromatin structure and other epigenetic factors such as histone modifications, DNA methylation and so on. Research in this field has witnessed a major leap forward with the introduction of techniques like Chromatin Immunoprecipitation (ChIP) (1), which, followed by the employment of genome tiling oligonucleotide arrays [ChIP on Chip (2)] or next-generation sequencing [ChIP-Seq (3)], permits to build genome-wide maps of TF binding, histone modifications or any other DNA interacting protein involved in transcription regulation. ChIP-Seq has rapidly become the method of choice for research in this field, given its higher resolution with respect to ChIP on Chip, and the constantly decreasing cost of next-generation sequencing experiments. As a consequence, today genomic resources like the UCSC Genome Browser (4) or dedicated databases like hmChip (5) make available for retrieval the genomic maps of hundreds of TFs, as well as of histone modifications, PolII and PolIII binding, and so on, in several different cell lines. Thus, starting from a gene, its putative regulators as well as epigenetic information associated with it can be easily retrieved vice versa, different ChIP-Seq experiments can be correlated with one another by comparing the distribution of the sequence reads of each one (6) and, once the list of genomic binding regions for a TF is available, the target genes it is likely to regulate can be easily singled out by using tools like GREAT (7).

On the other hand, a very common problem that researchers have to face is, given a set of genes showing similar expression patterns, to find out which common regulators they share, responsible for the expression observed. This type of analysis is often performed by finding similar and over-represented sequence elements, for example in promoter sequences, either by using de novo motif finding tools (8) or descriptors of the binding specificity of TFs (9,10). While useful in many cases, these approaches suffer from several limitations: the binding

*To whom correspondence should be addressed. Tel: +39 50314884; Fax: +39 50315044; Email: giulio.pavesi@unimi.it specificity of many TFs is as yet unknown or not well characterized; different TFs have very similar binding sites, making difficult, given a sequence motif, to assess which TF actually could bind it; some key regulators do not bind DNA directly, but act as co-factors with TFs; sequence motifs are often weakly conserved, and hard to discriminate against random similarities; sequence analysis tools usually ignore chromatin structure and DNA accessibility, usually resulting in an 'over-prediction' of sequence motifs.

The web tool we present, named Cscan, is based on a large collection of ChIP-Seq experiments for several TFs and other factors related to transcription regulation. Enriched regions from the ChIP-Seq experiments have been crossed with the genomic coordinates of available RefSeq and Ensembl gene annotations, so to build genome-wide maps of putative target genes in each experiment. Given a set of genes as input, the interface evaluates the over- (or under-) representation of target sites for the DNA binding protein considered in each ChIP experiment by counting the number of target genes in the experiment contained in the input set, and comparing this count to the overall genome-wide number of its target genes to assess statistical significance with a Fisher's exact test. Experiments with a significantly high number of sites within the input genes' loci are thus likely to correspond to TFs, which are common regulators of the genes. The computation is performed for hundreds of different TFs with other data like histone modifications and RNA polymerases (and/or their subunits), so to provide a more comprehensive view of all the genetic and epigenetic factors involved in the regulation of the input genes, and their effect on gene transcription.

ChIP-SEQ DATA COLLECTION

We retrieved the ChIP-Seq peak lists publicly available and already past the public release date at the UCSC Genome Browser for TFs, histone modifications, and RNA polymerases produced by the ENCODE project (11). Also, we retrieved from the original publications the datasets that have been included in the hmChip database (5). Finally, we added the HMMChip tracks of the UCSC Genome Browser, showing chromatin state segmentation for each of nine human cell types. A common set of states (including for example active promoter, weak promoter, repressed transcription, and so on) across the cell types were annotated integrating ChIP-Seq data for nine histone modifications using a Hidden Markov Model. The genome was thus segmented into regions according to the corresponding chromatin state (12).

Overall, data collection resulted for human in 409 different experiments for 144 TFs or co-factors in 65 different cell lines, 234 experiments for 11 different histone modifications in 23 cell lines, 46 experiments for 6 RNA polymerases (or their subunits) in 28 cell lines, data for CTCF binding in 49 cell lines, for a total of 777 different experiments or annotations in 102 cell lines. We are currently populating the mouse collection, which as of today contains data for about 50 TFs.

In each ChIP-Seq dataset, the genomic coordinates of each region marked as 'peak' have been crossed with the RefSeq or Ensembl gene annotations available. This resulted in a table with one row for each annotated gene, and one column for each ChIP-Seq experiment. The table reports the presence/absence of a peak in the ChIP-Seq experiment within different regions of the locus of the gene (i.e. in its promoter/upstream of the TSS at different distances, within the transcribed region, and so on, see Supplementary Figure S1).

FINDING COMMON REGULATORS

Starting from the data collected, let G be a sample of k genes or transcripts. If a given TF is a common regulator of the genes, then one should find an enrichment of binding regions for the TF associated with the genes, e.g. in their promoters or transcribed regions. For example, let m be the number of genes in the sample that have a peak for the TF in their promoter. Then, let N be the number of annotated genes in the genome and let n be the number of annotated genes in the genome that contain a ChIP-Seq peak for the TF in their promoter. The enrichment of the TF binding sites with respect to the gene sample can be thus evaluated by using a Fisher's exact test (hypergeometric distribution) with N, n, k and m as parameters.

The same principle can be applied to any other type of genome-wide ChIP experiment. For example, we can assess whether a given histone modification can be associated with the genes' promoters, hence denoting e.g. if their promoters as active or repressed, or whether RNA polymerase binding in a given cell line is enriched in the set of genes denoting their transcription, and so on.

THE USER INTERFACE

The user interface contains two main panels on the left and right hand side, which can be used to input a set of genes for finding their common regulators or for browsing and retrieving data from the ChIP-Seq data collections available.

Data browsing and retrieval

The right hand panel allows users to browse the data Cscan is based on, and to retrieve the list of target genes associated with a given experiment of interest. Users can select (i) the protein that has been ChIP'ed (ii) the cell line in which the experiment was performed, (iii) the region of the genes' locus in which peaks have to be located for the gene to be considered as a target (e.g. the −450, +50 region including the core promoter or the transcribed region including 1 kbp upstream), (iv) the source organism and the gene annotation to be used to display the results (RefSeq or Ensembl) and (v) the genome assembly used in the study. Once any of the input fields is selected, the other choices are automatically limited to the experiments available, e.g. once a TF has been selected in the list, the selection of the cell lines will be limited to those for which data are available for the TF, and so on.

The output will be displayed within an 'Experiment view' output window, described in the 'Output' section. Alternatively, given a gene (transcript) identifier, users can retrieve the list of ChIP-Seq experiments in the database that present a peak within the gene region defined.

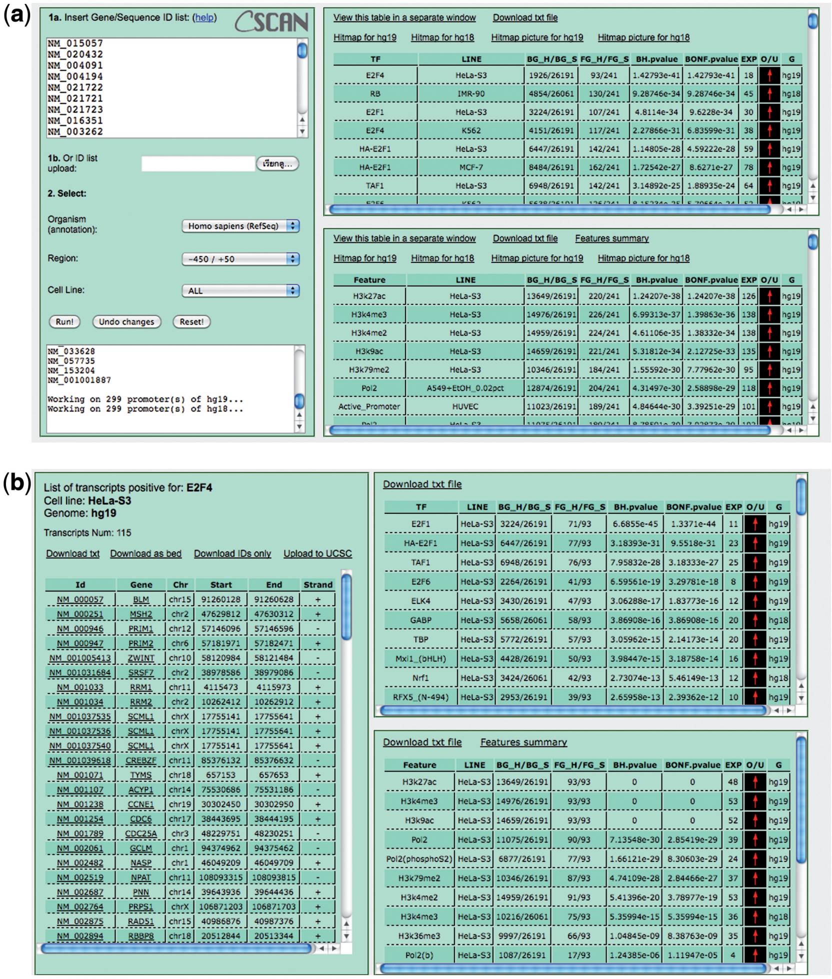
Gene input

The left hand panel is used to input a set of genes, by using the RefSeq or Ensembl IDs of their respective transcripts, and finding ChIP-Seq experiments that have a significantly high (or low) number of peaks associated with the genes. Users then have to specify the following: (i) the source organism of the genes (at the present time, human or mouse); (ii) the region, with respect to the gene, that has to be analyzed (e.g. core promoter only or upstream and transcribed regions); and (iii) the cell line in which the ChIP-Seq data used for the analysis were generated (or this parameter can be set to 'ALL' indicating that all the data available have to be used). A typical analysis takes a few seconds, and results will appear in the middle of the page.

Output

The output is split into two tables as shown in Figure 1a. The topmost one is dedicated to TFs (or co-factors), while the bottom one contains results for CTCF, histone modifications, PolII and PolIII binding, HMMChip regions, and other experiments not involving TFs (denoted as 'Features' in the table). A link on the top of the features table gives further explanations on each one, and its possible effect on the regulation of the genes. In each table, the ChIP-Seq experiments used in the analysis are ranked according to the *P*-value of the Fisher's test. From left to right the columns of the output table summarize the following:

- [TF/FEATURE]: The TF or feature of the ChIP-Seq experiment;
- [LINE]: The cell line in which the experiment was performed;



1a. Insert Gene/Sequence ID list: (help)

```
NM_015057
NM_020432
NM_004091
NM_004194
NM_021722
NM_021721
NM_021723
NM_016351
NM_003262
```

1b. Or ID list upload:

2. Select:

Organism (annotation): Homo sapiens (RefSeq)

Region: -450 / +50

Cell Line: ALL

```
NM_033628
NM_057735
NM_153204
NM_001001887

Working on 299 promoter(s) of hg19...
Working on 299 promoter(s) of hg18...
```

View this table in a separate window   Download txt file

Hitmap for hg19   Hitmap for hg18   Hitmap picture for hg19   Hitmap picture for hg18

TF	LINE	BG_H/BG_S	FG_H/FG_S	BH.pvalue	BONF.pvalue	EXP	O/U	G
E2F4	HeLa-S3	1926/26191	93/241	1.42793e-41	1.42793e-41	18	↑	hg19
RB	IMR-90	4854/26061	130/241	9.28746e-34	9.28746e-34	45	↑	hg18
E2F1	HeLa-S3	3224/26191	107/241	4.8114e-34	9.6228e-34	30	↑	hg19
E2F4	K562	4151/26191	117/241	2.27866e-31	6.83599e-31	38	↑	hg19
HA-E2F1	HeLa-S3	6447/26191	142/241	1.14805e-28	4.59222e-28	59	↑	hg19
HA-E2F1	MCF-7	8484/26191	162/241	1.72542e-27	8.6271e-27	78	↑	hg19
TAF1	HeLa-S3	6948/26191	142/241	3.14892e-25	1.88935e-24	64	↑	hg19

View this table in a separate window   Download txt file   Features summary

Hitmap for hg19   Hitmap for hg18   Hitmap picture for hg19   Hitmap picture for hg18

Feature	LINE	BG_H/BG_S	FG_H/FG_S	BH.pvalue	BONF.pvalue	EXP	O/U	G
H3k27ac	HeLa-S3	13649/26191	220/241	1.24207e-38	1.24207e-38	126	↑	hg19
H3k4me3	HeLa-S3	14976/26191	226/241	6.99313e-37	1.39863e-36	138	↑	hg19
H3k4me2	HeLa-S3	14959/26191	224/241	4.61106e-35	1.38332e-34	138	↑	hg19
H3k9ac	HeLa-S3	14659/26191	221/241	5.31812e-34	2.12725e-33	135	↑	hg19
H3k79me2	HeLa-S3	10346/26191	184/241	1.55592e-30	7.77962e-30	95	↑	hg19
Pol2	A549+EtOH_0.02pct	12874/26191	204/241	4.31497e-30	2.58898e-29	118	↑	hg19
Active_Promoter	HUVEC	11023/26191	189/241	4.84644e-30	3.39251e-29	101	↑	hg19

List of transcripts positive for: **E2F4**
Cell line: **HeLa-S3**
Genome: **hg19**

Transcripts Num: 115

Download txt   Download as bed   Download IDs only   Upload to UCSC

Id	Gene	Chr	Start	End	Strand
NM_000057	BLM	chr15	91260128	91260628	+
NM_000251	MSH2	chr2	47629812	47630312	+
NM_000946	PRIM1	chr12	57146096	57146596	-
NM_000947	PRIM2	chr6	57181971	57182471	+
NM_001005413	ZWINT	chr10	58120984	58121484	-
NM_001031684	SRSF7	chr2	38978586	38979086	-
NM_001033	RRM1	chr11	4115473	4115973	+
NM_001034	RRM2	chr2	10262412	10262912	+
NM_001037535	SCML1	chrX	17755141	17755641	+
NM_001037536	SCML1	chrX	17755141	17755641	+
NM_001037540	SCML1	chrX	17755141	17755641	+
NM_001039618	CREBZF	chr11	85376132	85376632	-
NM_001071	TYMS	chr18	657153	657653	+
NM_001107	ACYP1	chr14	75530686	75531186	-
NM_001238	CCNE1	chr19	30302450	30302950	+
NM_001254	CDC6	chr17	38443695	38444195	+
NM_001789	CDC25A	chr3	48229751	48230251	-
NM_002061	GCLM	chr1	94374962	94375462	-
NM_002482	NASP	chr1	46049209	46049709	+
NM_002519	NPAT	chr11	108093315	108093815	-
NM_002687	PNN	chr14	39643936	39644436	+
NM_002764	PRPS1	chrX	106871203	106871703	+
NM_002875	RAD51	chr15	40986876	40987376	+
NM_002894	RBBP8	chr18	20512844	20513344	+

Download txt file

TF	LINE	BG_H/BG_S	FG_H/FG_S	BH.pvalue	BONF.pvalue	EXP	O/U	G
E2F1	HeLa-S3	3224/26191	71/93	6.6855e-45	1.3371e-44	11	↑	hg19
HA-E2F1	HeLa-S3	6447/26191	77/93	3.18393e-31	9.5518e-31	23	↑	hg19
TAF1	HeLa-S3	6948/26191	76/93	7.95832e-28	3.18333e-27	25	↑	hg19
E2F6	HeLa-S3	2264/26191	41/93	6.59561e-19	3.29781e-18	8	↑	hg19
ELK4	HeLa-S3	3430/26191	47/93	3.06288e-17	1.83773e-16	12	↑	hg19
GABP	HeLa-S3	5658/26061	58/93	3.86908e-16	3.86908e-16	20	↑	hg18
TBP	HeLa-S3	5772/26191	57/93	3.05962e-15	2.14173e-14	20	↑	hg19
Mxi1_(bHLH)	HeLa-S3	4428/26191	50/93	3.98447e-15	3.18758e-14	16	↑	hg19
Nrf1	HeLa-S3	3424/26061	42/93	2.73074e-13	5.46149e-13	12	↑	hg18
RFX5_(N-494)	HeLa-S3	2953/26191	39/93	2.65958e-13	2.39362e-12	10	↑	hg19

Download txt file   Features summary

Feature	LINE	BG_H/BG_S	FG_H/FG_S	BH.pvalue	BONF.pvalue	EXP	O/U	G
H3k27ac	HeLa-S3	13649/26191	93/93	0	0	48	↑	hg19
H3k4me3	HeLa-S3	14976/26191	93/93	0	0	53	↑	hg19
H3k9ac	HeLa-S3	14659/26191	93/93	0	0	52	↑	hg19
Pol2	HeLa-S3	11075/26191	90/93	7.13548e-30	2.85419e-29	39	↑	hg19
Pol2(phosphoS2)	HeLa-S3	6877/26191	77/93	1.66121e-29	8.30603e-29	24	↑	hg19
H3k79me2	HeLa-S3	10346/26191	87/93	4.74109e-28	2.84466e-27	37	↑	hg19
H3k4me2	HeLa-S3	14959/26191	91/93	5.41396e-20	3.78977e-19	53	↑	hg19
H3k4me3	HeLa-S3	10216/26061	75/93	5.35994e-15	5.35994e-15	36	↑	hg18
H3k36me3	HeLa-S3	9997/26191	66/93	1.04845e-09	8.38763e-09	35	↑	hg19
Pol2(b)	HeLa-S3	1087/26191	17/93	1.24385e-06	1.11947e-05	4	↑	hg19

Figure 1. (**a**) Example of the output of Cscan showing the list of input genes (input box on the left hand side) as well as the TFs list (right, top) and 'features' (histone modifications, polymerase binding and so on, bottom right) ranked according to the resulting *P* value. See the main text for further explanation on the output fields. (**b**) The 'Experiment view' table, showing for a TF (E2F4) in the selected cell line (HeLa-S3) the target genes that were included in the input sample (left). The tables on the right show enrichment of other TFs (top) and features (bottom), computed on the E2F4 target genes.

- [BG_H/BG_S]: The number BG_H of genes in the genome annotation which contain a peak for the experiment in the region selected as input, and the overall number BG_S of 'background' genes (e.g. in the whole genome annotation used);
- [FG_H/FG_S]: As in the previous point, but restricted to the FG_H genes that contain a peak out of the total FG_S input genes;
- [Bonf-Pvalue]: The Bonferroni corrected *P* value computed with the Fisher's test (hypergeometric distribution) according to the BG_S, BG_H, FG_S and FG_H values;
- [BH-Pvalue]: The Benjamini-Hochberg corrected *P* value. This correction yields less restrictive *P* values than the Bonferroni one. Users can choose which of the two seems to be more suitable for their data;
- [EXP]: The expected value for FG_H, according to FG_S, BG_H and BG_S; and
- [O/U]: A red arrow pointing upwards if the number of FG_H genes is greater than the expected value, a green arrow pointing downwards if lower. The arrow thus denotes whether peaks for the ChIP-Seq experiment are under- or over-represented in the gene sample provided as input.

Experiments that present a large number of genome-wide targets (i.e. more than one third of the annotated transcripts), thus unlikely to provide meaningful information, have the corresponding line shaded in grey. Clicking on the links above each table allows users to open the table in a new window (discussed later), to download the table in text format, or to display the 'hitmap', which associates with each input gene and ChIP experiment pair a '1' if the gene region specified as input contains at least one ChIP peak, '0' if not, and that can be used for further computations and analyses. The hitmap can be displayed also as a picture, with a colored spot in correspondence of experiment peak-gene associations, black otherwise.

Once an output table is opened in a new window (called 'TF table view' and 'Feature view', respectively), additional links appear. Clicking on the 'info' icons (a white 'i' on a blue background) provides further information (if available) on the subject of the ChIP experiment or the cell line in which it was performed. Clicking on the TF/feature name, instead, opens the 'Experiment view' window, that displays the list of input genes which are associated with a peak for the ChIP-Seq experiment and cell line selected, as well as their genomic coordinates. From this window, users can download the list of gene IDs, or the .bed file of their genomic coordinates which can be uploaded automatically to the UCSC Genome Browser for further analysis. The 'Get Correlations' button on the right-hand side performs another run of Cscan, but restricted to the list of target genes for the TF/feature currently investigated, and using experiments performed in the same cell line: in this way, users can immediately assess which other experiments have significant correlation (or anti-correlation) with the TF/feature on the set of genes studied (Figure 1b).

This 'Experiment view' window is also displayed when a given experiment is selected by using the right-hand panel of the interface: in this case, the list of targets will comprise all the target genes in the genome annotation available.

EXAMPLES

As mentioned before, Cscan can be applied to different types of analysis. A straightforward application is to study clusters of genes with similar expression patterns, so to single out their putative common regulators. But, as epigenetic data are also included in the analysis, by crossing these data with the results on TF binding one can get an idea also on the effect of the TF regulation (activation/repression) and/or on the tissue/cell/condition specificity of TF binding. Also, if a novel ChIP-Seq experiment has been performed, Cscan allows for an immediate assessment of which other TFs show significance correlation or anti-correlation with the studied one, as well as of whether the TF correlates with histone modifications, active/repressed promoters, or polymerase binding, the latter indicating again whether it might act as an activator or a repressor. Finally, the results of Cscan provide an immediate validation for predictions derived from other tools, for example conserved motifs found by sequence analysis and motif discovery methods. Users can thus submit the same set of genes simultaneously to Cscan and to tools like Pscan (9) or Clover (10), and assess whether TFs singled out by sequence analysis are also detected by Cscan, or which TFs are more likely to bind a given sequence motif. These two approaches can also be seen as complementary, because ChIP data are not available for all the TFs and vice versa, a reliable binding descriptor is not available for all the TFs.

In the following section, we describe some examples of usage of Cscan. The corresponding datasets are included in the interface and can be easily loaded as input by clicking on the corresponding link. The analyses were performed by using as a target region of the genes the core promoter (from −450 to +50 with respect to the transcription start site). The results are also provided in Supplementary Table S1.

Human cell cycle regulated genes

We retrieved the clusters of human genes whose expression has been characterized of being specific of a given phase of the cell cycle [G1/S, S, G2, G2/M and M/G1 (13)]. The microarray experiment was performed in HeLa cells.

Concerning the 'features' table, nearly all the genes of each dataset seem to be transcribed in all the cell lines available, and not only HeLa cells. Indeed, PolII, 'Active promoter' HMMChIP annotations, and histone modifications associated with active promoters and transcription are highly enriched, while those features associated with gene silencing are significantly under represented. This is hardly a surprising result, because we can expect cell-cycle expressed genes not to be cell line or tissue specific. Figure 2 and Supplementary Table 1 summarize the most significant TFs in the five phases (Bonferroni corrected $P < 10^{-5}$ in at least one).

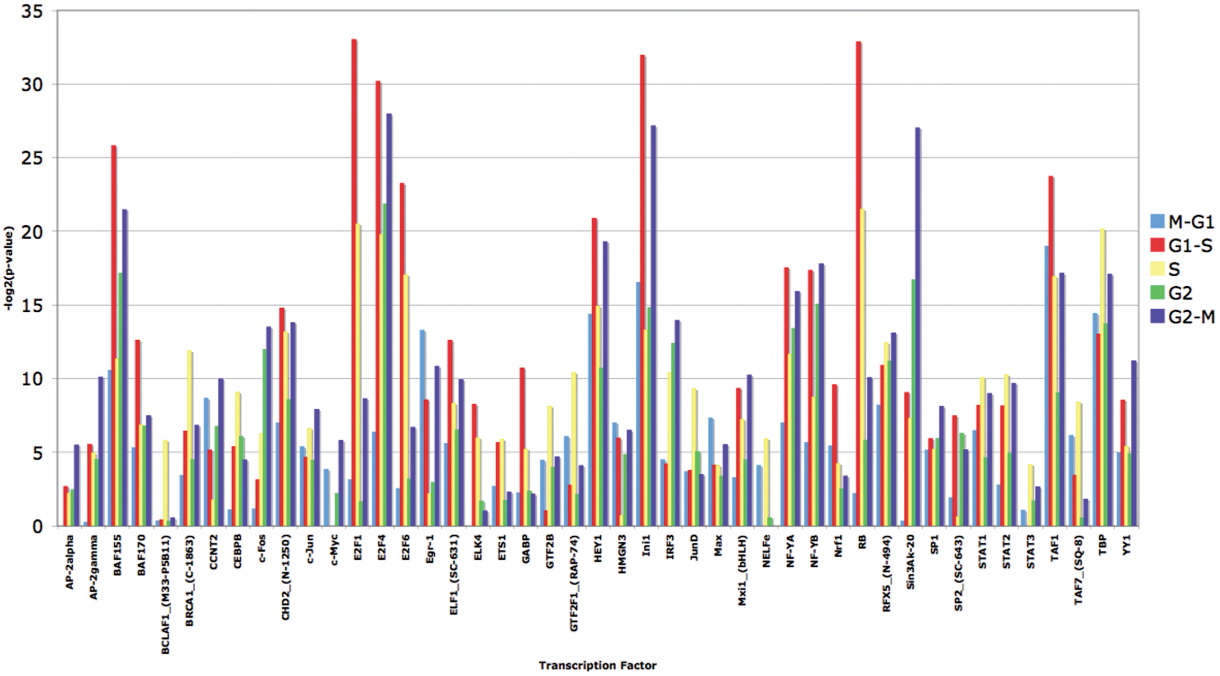


Figure 2. The most significantly enriched TFs in the different phases of human cell cycle ($P < 10^{-5}$ in at least one set). We considered experiments performed on the same cell line of the expression data (HeLa). For TFs for which HeLa data are not yet available, we employed K562 data.

TFs of the list showing highest enrichments are known cell cycle regulators, and as expected their over-representation in the input sets changes according to the different phases. For example, all members of the E2F family with available data are significantly enriched in phases G1/S and S. E2F1 and E2F6, however, drop in the successive phases, while E2F4 remains enriched throughout G2 and G2/M. The Retinoblastoma protein (RB) lacks a DNA binding domain and is recruited to promoters by other sequence-specific TFs, such as the members of the E2F family: indeed its enrichment profile shows similarity with E2F1 and E2F6. Thus, results of this kind would be hard to obtain with sequence analysis alone, also because specialized databases like TRANSFAC and JASPAR report only a generic 'E2F' binding motif, while for example E2F4 has been recently shown to bind the CHR promoter element as part of the DREAM complex (14).

Human tissue-specific genes

As an example of analysis of tissue-specific genes, we built two datasets of liver and whole brain-specific genes using the Gene Sorter tool at the UCSC Genome Browser. We selected RefSeq genes with an expression logBase2 (tissue/reference) value >2 in the two tissues.

The 'features' results on liver genes show how their transcription activation, active promoter marks and PolII binding seem to be confined to HepG2 cells, which indeed are hepatocarcinoma cells, and a model system for the study of polarized human hepatocytes. On the other hand, the signatures associated to transcription repression and gene silencing are over-represented in all other cell lines. Also, the TF table shows as significantly enriched a series of TFs (HNF4A, HNF4G, RXRA, FOXA1, and so on) known to be associated with tissue-specific gene expression in liver or liver development. Other TFs usually associated with cell cycle or 'housekeeping' gene expression do not show any enrichment on this gene set in HepG2 cells. However, not all the genes of this set are associated with PolII binding or active promoters and TF binding. This fact can be due to different reasons, like experimental issues (false positives in the microarray experiment or false negatives in the ChIP-Seq analyses producing the lists of peaks), or to differences between normal and tumoral liver cells. Another possibility is that, as multiple promoters can be associated with the same gene, Cscan is able to mark which promoter is active and bound by TFs in the cell line investigated.

The result on 'brain-specific' genes shows how they do not seem to be transcribed in any of the cell lines available, nor are enriched for any histone mark associated with transcriptional activation. The TF table likewise shows how virtually all the TFs are underrepresented in the gene sample, with the sole exception of NRSF (Neuron-restrictive Silencer transcription Factor) throughout different cell lines, which indeed is a repressor protein expressed in non-neuronal tissues, repressing the expression of several neuronal genes.

Computing correlations between different ChIP-Seq experiments

A simple but explicative example on how Cscan can be used to identify correlations among different ChIP-Seq experiments is the set of target promoters for BDP1 (B double-prime 1) in human HeLa cells, retrievable from Cscan itself. BDP1 is a subunit of the TFIIIB transcription initiation complex, which recruits RNA polymerase III to target promoters to activate its transcription (15). Indeed, the features table shows that PolIII associated with the promoters of the genes. In the TF list, the highest correlations are with BDP1 itself in a different cell line (K562), showing how BDF1 binding does not seem to be cell-line specific. Also, all the targets of another factor (BRF1) are included into the BDF1 list. Indeed, BDF1 is another subunit of the same complex, together with TFIIIC-110, which is also highly enriched. Other regulators, not related to PolIII transcription appear anyway to be over-represented. Although, for example TATA-binding protein has already been shown to be a regulator of PolIII transcribed genes, other factors like STAT1 in interferon-stimulated cells or heat-shock protein that target most of the genes show how they are probably activated and involved in several different pathways.

CONCLUSIONS

Cscan is a web server that employs a collection of several hundreds of different ChIP-Seq experiments to identify putative common regulators in a set of genes, as well as assessing their transcriptional and epigenetic profile. Clearly, results depend on the presence of a given TF or cell line in the collection of experiments the server is based on, and while for example we have already a good coverage for tissues like liver we still lack data on tissue-specific TFs and epigenomic information in several tissues or cell lines. We can expect however this gap to be quickly filled in the near future, given the ever increasing amount of ChIP-Seq experiments that are performed and published almost on a daily basis. Also, we plan in the near future to include information about distal regulatory elements in enhancers/silencers, by crossing data on TF binding with chromatin signatures marking likely enhancer regions and with CTCF binding to insulators, so to overcome the obvious limitations of analyses only on promoters or transcribed regions. Concerning the extension of Cscan to other species, we are currently populating the ChIP-Seq mouse data collection, as well as preparing the inclusion of other species, from yeast to the data produced by the modENCODE project on *Caenorhabditis elegans* and *Drosophila* (16).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figure 1.

FUNDING

Italian Ministry of University and Research Fondo Italiano per la Ricerca di Base (FIRB) project 'Laboratorio Internazionale di Bioinformatica' (LIBI); Consiglio Nazionale delle Ricerche (CNR) flagship project EPIGEN. Funding for open access charge: Italian Ministry of University and Research Fondo Italiano per la Ricerca di Base (FIRB) project 'Laboratorio Internazionale di Bioinformatica' (LIBI).

Conflict of interest statement. None declared.

REFERENCES

1. Collas,P. and Dahl,J.A. (2008) Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front. Biosci.*, **13**, 929–943.
2. Pillai,S. and Chellappan,S.P. (2009) ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. *Methods Mol. Biol.*, **523**, 341–366.
3. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
4. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
5. Chen,L., Wu,G. and Ji,H. (2010) hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics*, **27**, 1447–1448.
6. Ye,T., Krebs,A.R., Choukrallah,M.A., Keime,C., Plewniak,F., Davidson,I. and Tora,L. (2011) seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.*, **39**, e35.
7. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
8. Pavesi,G., Mauri,G. and Pesole,G. (2004) In silico representation and discovery of transcription factor binding sites. *Brief Bioinform.*, **5**, 217–236.
9. Zambelli,F., Pesole,G. and Pavesi,G. (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.*, **37**, W247–W252.
10. Frith,M.C., Fu,Y., Yu,L., Chen,J.F., Hansen,U. and Weng,Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.
11. Rosenbloom,K.R., Dreszer,T.R., Long,J.C., Malladi,V.S., Sloan,C.A., Raney,B.J., Cline,M.S., Karolchik,D., Barber,G.P., Clawson,H. *et al.* (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.*, **40**, D912–D917.
12. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
13. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell.*, **13**, 1977–2000.
14. Muller,G.A., Quaas,M., Schumann,M., Krause,E., Padi,M., Fischer,M., Litovchick,L., Decaprio,J.A. and Engeland,K. (2012) The CHR promoter element controls cell cycle-dependent gene transcription and binds the DREAM and MMB complexes. *Nucleic Acids Res.*, **40**, 1561–1578.
15. Noma,K. and Kamakaka,R.T. (2010) The human Pol III transcriptome and gene information flow. *Nat. Struct. Mol. Biol.*, **17**, 539–541.
16. Contrino,S., Smith,R.N., Butano,D., Carr,A., Hu,F., Lyne,R., Rutherford,K., Kalderimis,A., Sullivan,J., Carbon,S. *et al.* (2012) modMine: flexible access to modENCODE data. *Nucleic Acids Res.*, **40**, D1082–D1088.

Genome **Biology**

RESEARCH PAPER **Open Access**

Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in *Arabidopsis*

Veronica Gregis[1†], Fernando Andrés[2†], Alice Sessa[1†], Rosalinda F Guerra[1], Sara Simonini[1], Julieta L Mateos[2], Stefano Torti[2], Federico Zambelli[1], Gian Marco Prazzoli[1], Katrine N Bjerkan[3], Paul E Grini[3], Giulio Pavesi[1], Lucia Colombo[1,4], George Coupland[2] and Martin M Kater[1*]

Abstract

Background: MADS-domain transcription factors play important roles during plant development. The *Arabidopsis* MADS-box gene *SHORT VEGETATIVE PHASE* (*SVP*) is a key regulator of two developmental phases. It functions as a repressor of the floral transition during the vegetative phase and later it contributes to the specification of floral meristems. How these distinct activities are conferred by a single transcription factor is unclear, but interactions with other MADS domain proteins which specify binding to different genomic regions is likely one mechanism.

Results: To compare the genome-wide DNA binding profile of SVP during vegetative and reproductive development we performed ChIP-seq analyses. These ChIP-seq data were combined with tiling array expression analysis, induction experiments and qRT-PCR to identify biologically relevant binding sites. In addition, we compared genome-wide target genes of SVP with those published for the MADS domain transcription factors FLC and AP1, which interact with SVP during the vegetative and reproductive phases, respectively.

Conclusions: Our analyses resulted in the identification of pathways that are regulated by SVP including those controlling meristem development during vegetative growth and flower development whereas floral transition pathways and hormonal signaling were regulated predominantly during the vegetative phase. Thus, SVP regulates many developmental pathways, some of which are common to both of its developmental roles whereas others are specific to only one of them.

Keywords: MADS-box, gene regulation, transcription factors, post transcriptional regulation, ChIP-seq, floral transition, floral development, *Arabidopsis thaliana*

Background

In plants organs are formed post-embryonically from populations of undifferentiated cells called meristems. In these meristems, stem cell activity is kept at the central zone whereas at the peripheral part of the meristem primordia arise in which cells differentiate into organs. In flowering plants like *Arabidopsis thaliana* during the vegetative phase the primordia that derive from the shoot apical meristem (SAM) develop into leaves [1,2]. The change to the subsequent generative phase is called floral transition, which is regulated by multiple flowering pathways that are controlled by environmental and endogenous cues. During the floral transition, the SAM undergoes a change in fate and becomes an inflorescence meristem (IM). The *Arabidopsis* IM is an indeterminate meristem and develops multiple determinate floral meristems (FMs) in a spiral manner, which in turn produce a precise number of floral organs arranged in a whorled pattern [1,3,4]. The reprogramming of meristems is regulated by a complex gene regulatory network in which transcription factors represent important key players.

* Correspondence: martin.kater@unimi.it
† Contributed equally
[1]Department of Bioscience, Università degli Studi di Milano, Via Celoria 26, 20133 Milan, Italy
Full list of author information is available at the end of the article

In Arabidopsis the photoperiod, thermosensory, and vernalization/autonomous pathways that respond to environmental signals, and the aging and gibberellic acid pathways that respond to the developmental and physiological state of the plant regulate the floral transition [5]. Many transcription factors encoding genes have been shown to be involved in the regulation of these pathways including those belonging to the MADS-box gene family [6,7]. One of these MADS-box genes controlling flowering time is *SHORT VEGETATIVE PHASE* (*SVP*) [8].

MADS-domain transcription factors have been identified in all eukaryotic kingdoms and in *Arabidopsis thaliana* they are involved in most important developmental processes [9-12]. MADS-domain factors activate or repress transcription by direct binding to short sequences called CArG-boxes that correspond to a 10 nucleotide sequence $CC(A/T)_6GG$ present in the regulatory sequences of target genes. However, this motif can be quite variable allowing some mismatches [10,13]. Moreover MADS-domain proteins form homo and/or heterodimers and are also suggested to form tetrameric MADS-domain complexes [14]. The variety of interactions that many MADS-domain factors can make suggests that they may regulate different subsets of genes during different phases of development and might reflect an enormous regulatory potential [15]. Furthermore, their association with others co-factors probably also influences the affinity and specificity of the complex for specific target sequences [16,17].

During the vegetative phase *SVP* acts as a repressor of flowering since the *svp* mutant flowers very early [8]. *SVP* mediates flowering responses by perceiving signals from different endogenous and environmental flowering pathways such as the thermosensory, autonomous, and GA pathways [6,18]. SVP regulates the expression of three floral pathways integrator genes (FPI) that are *FLOWERING LOCUS T* (*FT*), *TWIN SISTER OF FT* (*TSF*), and *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) which all promote flowering [18,19]. To maintain plants in the vegetative phase, SVP represses the expression of *FT* and *TSF* in the phloem and *SOC1* in the SAM by directly binding to CArG boxes in *FT* and *SOC1* [6,18,19]. During the vegetative phase, SVP interacts with another central repressor of flowering time that is FLOWERING LOCUS C (FLC) and their function is mutually dependent. In fact it has recently been demonstrated that the SVP-FLC dimer acts to directly repress *FT* in the leaves and *SOC1* in the SAM [18]. During the floral transition, *SVP* expression gradually decreases until the SVP protein completely disappears from the IM [20]. In plants competent to flower, inputs deriving from the flowering pathways converge to repress SVP and FLC expression [18,19]. During the vegetative phase *SVP* plays an opposite role to its phylogenetically closest related MADS-box gene *AGAMOUS LIKE 24* (*AGL24*), which is a central promoter of flowering [21,22]. Both SVP and AGL24 directly regulate *SOC1* by binding its promoter on the same binding sites but they have an opposite effect on *SOC1* expression [23].

Interestingly, after the floral transition both *SVP* and *AGL24* are co-expressed in the floral meristem during stage 1 and 2 of flower development [24]. Analysis of the *svp agl24* double mutant, especially at higher temperatures, and the *svp ap1 agl24* triple mutants showed that *AGL24* and *SVP* play redundant roles during these early stages of flower development [20,24,25]. Combining the *svp agl24* double mutant with a weak *ap1* allele showed that AGL24 and SVP together with AP1 repress floral homeotic genes that control petal, stamen and carpel identity [25]. Protein interaction and genetic studies revealed that SVP and AGL24 are able to form dimers with AP1 and that this dimer is able to recruit the LEUNIG-SEUSS co-repressor complex [15,25]. Combining the *svp agl24* double mutant with a strong *ap1* allele showed that they are also controlling floral meristem identity since this triple mutant forms on the flanks of the IM new IMs instead of FMs resulting in a cauliflower like curd just as observed in the *ap1 cauliflower* (*cal*) double mutant [24,26]. Recently Simonini *et al.* [17] have shown that the co-repressor complex composed of LUG, SEU, and SVP is also able to repress the ovule identity gene *SEEDSTICK* (*STK*) in a complex together with BASIC PENTACYSTEINE transcription factors.

SVP is a key factor for *Arabidopsis* development and acts both during vegetative and reproductive phases where it plays different roles probably by interacting with different partners to regulate specific sets of target genes. Even though *SVP* is a gene of interest since its first characterization [8], still little is known about the mode of action and the network of genes controlled by this MADS-domain transcription factor. A powerful tool to study *in vivo* the genome-wide DNA-binding patterns of transcription factors is the ChIP-seq technology that consists in ultra-high throughput Solexa (Illumina) sequencing of DNA samples obtained by chromatin immunoprecipitation (ChIP). This technique has been used for a few years to identify direct target genes. At first for human transcription factors like NRSF, STAT1, PPARγ, and FOXA2 [27-30] and recently this technology has been reported for the identification in *Arabidopsis* of genome wide targets of different MADS-domain proteins such as, SEPALLATA3 (SEP3), AP1, FLC, and SOC1 [13,31-33] and another important transcriptional regulator such as AP2 [34]. Moreover genome wide binding site analysis is also possible using the ChIP on chip method, as was done for AGAMOUS LIKE 15 (AGL15), LEAFY (LFY), SVP, and SOC1 [35-37].

Here we report the use of the ChIP-seq approach to identify genome wide binding sites for SVP, during two distinct developmental phases: the vegetative and reproductive phase. This study allowed us to identify new pathways that are regulated by SVP in vegetative and reproductive tissues and to investigate genome-wide interaction dynamics of a transcription factor during different phases of development.

Results

Genome-wide mapping of SVP binding sites during vegetative and reproductive development

For genome-wide identification of the *in-vivo* binding sites of the SVP MADS-box transcription factor ChIP was performed followed by single end-read sequencing with the Solexa/Illumina GA platform. For the ChIP experiments Arabidopsis *svp* mutant plants expressing epitope tagged SVP were used [20]. The full genomic region of *SVP* including 3 kb upstream of the start codon was cloned as a C-terminal fusion with GREEN FLUORESCENT PROTEIN (GFP) [38]. Since SVP plays important roles during two distinct non-overlapping phases of development, namely the floral transition [8] and the early stages (stages 1 and 2) of flower development [20,24,25,39], studying the genome-wide binding sites of SVP provides an opportunity to compare the pathways directly regulated by SVP during these two developmental phases. Therefore vegetative phase material was harvested from 2-week-old seedlings grown under short-day conditions, whereas reproductive phase inflorescences with developing flowers of stage 1 to 11 [40] were harvested to analyze its targets during flower development.

Several independent ChIP experiments were performed. As control the same tissues were harvested from wild-type plants that did not express SVP-GFP. ChIP experiments that showed relatively high enrichment for known SVP binding regions (*FT* for the vegetative tissues and *AG* for reproductive tissues) were used to select samples for sequencing (see Additional data file 1, Figure S1) [6,20].

Distribution of SVP binding sites across the genome and within genes

For both vegetative and reproductive tissues as well as for the control, two independent ChIP reactions were sequenced. As in similar experiments [13,31], sequence reads obtained from duplicate experiments for each of the three samples were pooled. Only reads mapping to a unique position on the genome were considered for further analysis. This resulted in about 3 million uniquely mapped reads for the two experiments using inflorescence material, 5 million for experiments performed using vegetative material, and 6 million for control experiments (Additional data file 1, Table S1).

The regions enriched for binding sites were then identified with a strategy broadly similar to the one previously employed for SEP3 and AP1 [13,31], and implemented in the CSAR tool [41]. At a Bonferroni-corrected *P* value of 0.01 this resulted in about 13,000 regions in inflorescence tissues and 25,000 in seedlings, reduced to about 8,000 and 15,000, respectively, at threshold 10-4, and about 1,300 in both experiments at threshold 10-5 (see material and methods and Additional data file 2, Table S2). The overall distribution of SVP-binding sites across the genome in both tissues does not change significantly, and shows that 40% of the sites are located within the 3 Kb upstream of the gene, 27% in the transcribed region, whereas 4% are inside the 1 Kb downstream regions (Figure 1a). Regions falling within the transcribed regions tend to be located towards the 3' UTR/transcription termination (Figure 1b). A similar observation was made on the genome-wide distribution of SEP3 MADS-box protein binding sites [13]; moreover in Kaufmann *et al.* [31] they found that AP1 is able to bind the 3' region of *TERMINAL FLOWER 1* (*TFL1*) which is an important shoot identity gene [42]. *TFL1* 3' region is indeed required for proper *TFL1* expression. To confirm binding sites of SVP a set of target genes containing predicted binding sites at the 3' end was selected and analyzed in detail. This set included *AGL24*, *SEEDSTICK* (*STK*), *APETALA3* (*AP3*), and *FLOWERING LOCUS C* (*FLC*). As shown in Figure 1c, these genes show peaks of enrichment in the inflorescence ChIP-seq data near their 3'UTR regions and, for *STK* and *FLC*, these regions correspond to predicted SVP binding sites (3'UTR is indicated by the striped rectangle). The enrichments on the 3' UTR were analyzed in independent ChIP-qPCR assays confirming that binding at the 3'UTR is significant (Figure 1d).

Candidate target genes were then identified by associating each gene with an overall *P* value calculated from the product of the *P* values of the single binding regions located across the whole gene, encompassing the 3 kb upstream of the transcription start site to 1 kb downstream of the transcribed region. Thus, genes could be ranked according to the overall *P* values obtained. Starting from the ranked gene lists, we selected as high-confidence targets 2,982 genes in seedlings (with a cumulative gene *P* value < 1.26E-23) and 2,993 genes in inflorescences (cumulative gene *P* value <3.16E-15) (Additional data file 2, Table S2). The cut-offs on these lists were selected to maximize the number of known targets while excluding the maximum number of genes that were demonstrated to be false positives based on validations with ChIP-qPCR.

Binding motifs of the SVP protein

MADS-domain proteins are known to bind to different CArG box sequences, including the SRF-type (CC[A/T]6GG), the MEF2-type (C[A/T]8G), and other intermediate motifs (CC[A/T]7G/C[A/T]7GG) [10,43-46]. In order to assess the enrichment of CArG box motifs within the binding regions obtained from ChIP-seq, and to determine whether there is a preferred form of CArG box for

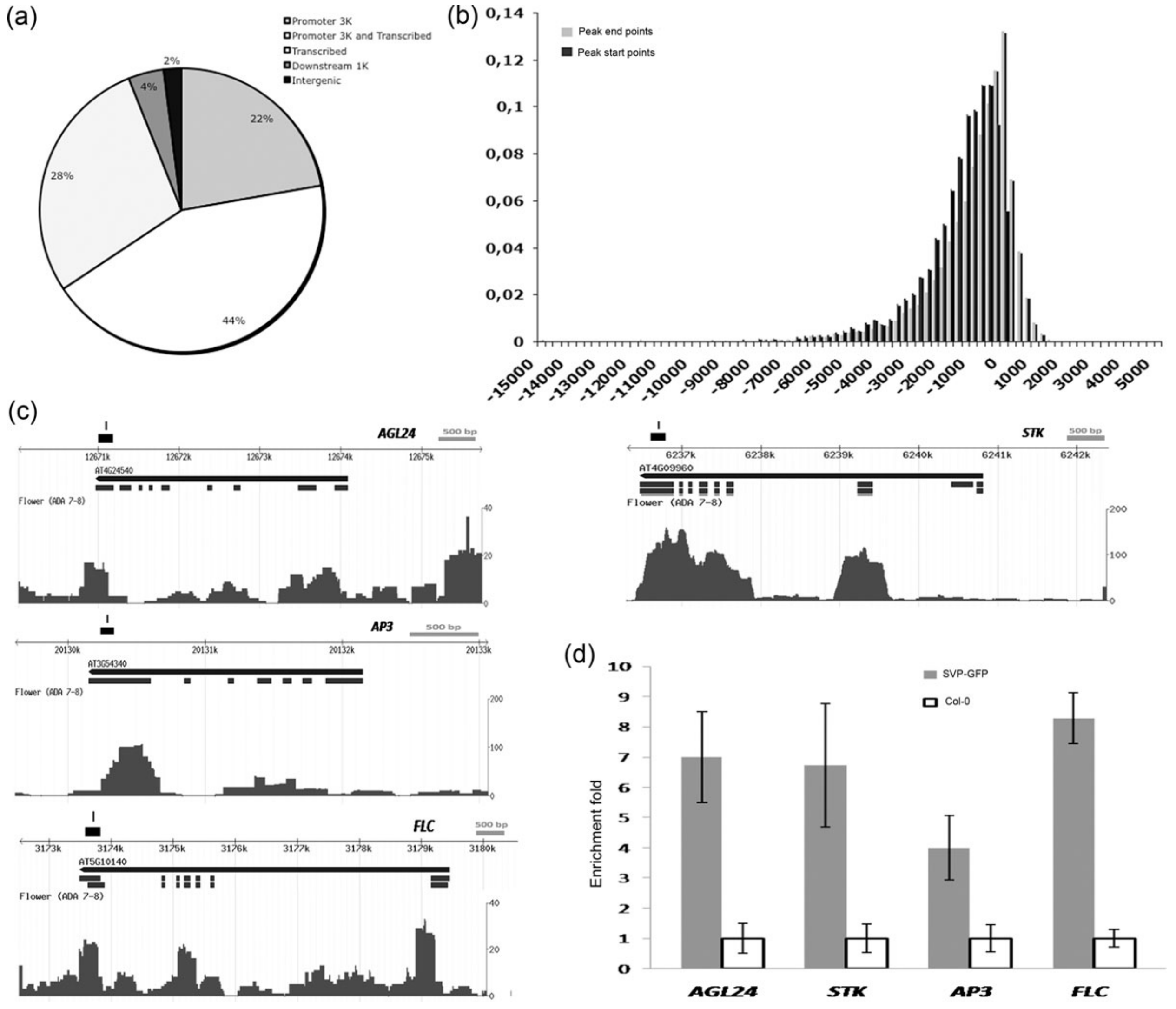


Figure 1 Location of SVP binding sites relative to nearby genes and analysis of SVP binding sites at the 3′ UTR regions of target genes. (**a**) Promoter 3K refers to the 3,000 bp upstream of the transcription start site (TSS); transcribed refers to the transcript from the 5′ UTR to 3′ UTR. Promoter 3K and Transcribed refers to 3,000 bp upstream to the TSS until the 3′UTR region. Downstream 1K starts from the transcription termination site until 1,000 bp downstream. Intergenic is none of the above regions; (**b**) diagram representing distribution of SVP binding (peaks) sites within the transcribed regions with respect to transcription termination sites (0 on the × axis); (**c**) binding profiles in inflorescence tissue for selected target genes which are bound by SVP in the transcribed regions: *AGL24*, *SEEDSTICK* (*STK*), *APETALA 3* (*AP3*), and *FLC*. TAIR annotation corresponds to TAIR8. Grey boxes represent the region validated by ChIP-PCR shown in (c); (**d**) ChIP-PCR validation for selected SVP target genes. ChIP assays were done using GFP antibodies and *SVP::SVP-GFP svp-41* plants and compared with wild-type control plants. Error bars represent standard deviations of normalized data (SD).

SVP, we ran a tailored version of the motif finder Weeder [47] in order to evaluate separately the enrichment within the regions of each oligonucleotide which could be considered a valid instance of a CArG box given the consensuses described before and also including NC[A/T]6GN. Oligonucleotides found to be enriched in the regions were then clustered together to form the motif maximizing the enrichment score. Motif enrichment was computed according to the Weeder score, which compares the number of occurrences within the ChIP enriched regions to an expected value derived from its number of occurrences genome-wide, computing a log ratio of the fold enrichment. The results are summarized in Figure 2a, split with respect to the two experiments performed and to the ranking of the ChIP regions according to their enrichment *P* value (best 1,000 regions, best 2,000, and so on). Enrichment clearly increases according to peak rank, with higher CArG box enrichment to be found within the peaks more enriched in the ChIP-seq experiments. Enrichment seems to be slightly higher in flower-enriched

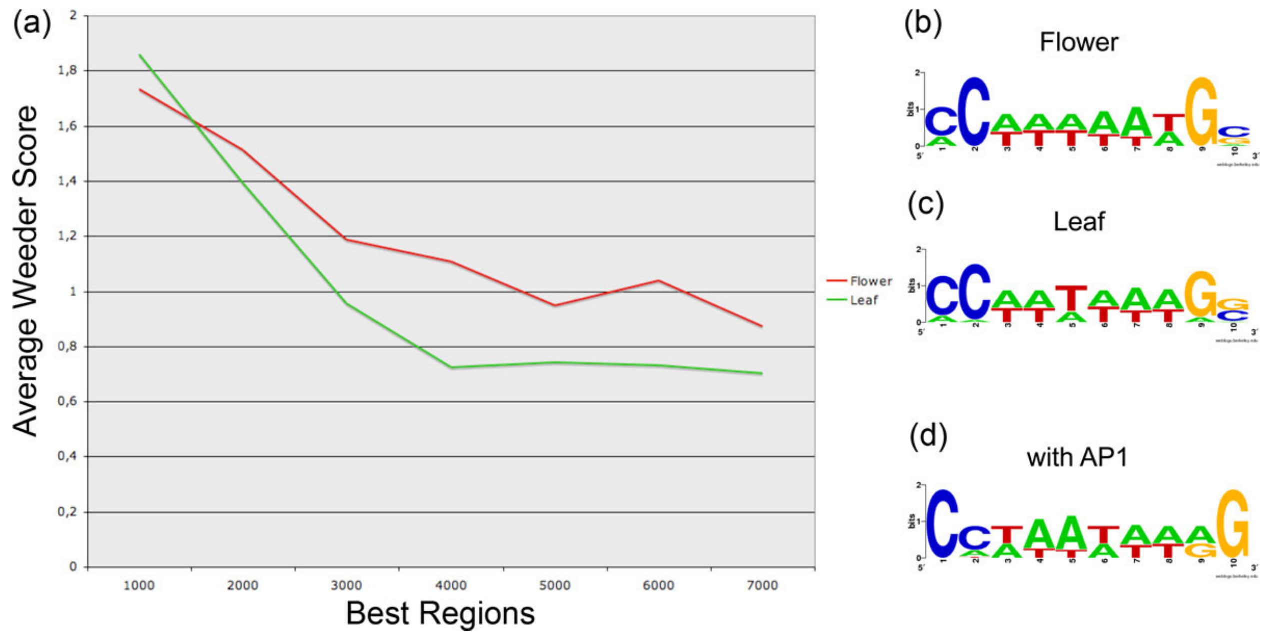


Figure 2 Enrichment of CArG box motifs within the binding regions obtained from ChIP-seq and CArG box for SVP. (**a**) Motif enrichment computed according to the Weeder score split with respect to the two experiments; (**b**) preferred consensus of most enriched oligos in flower; (**c**) preferred consensus of most enriched oligos in seedlings; (**d**) preferred consensus of most enriched oligos restricted to regions shared by SVP and AP1 in flowers.

regions with respect to leaf-enriched regions. Also, sequence alignment of most enriched oligos in flowers shows NC[A/T]6GN (shown in the sequence logo of Figure 2b and 2c) as a preferred consensus, which differs slightly from the already known forms briefly discussed above but closely resembles the one presented in Tao *et al*. [37]. Finally, oligo analysis restricted to regions shared by SVP and AP1 shows a more canonical CArG box, which is present in the regions with a much higher enrichment (about eight-fold enrichment with respect to the four-fold enrichment in the other regions; Figure 2d).

Comparison of SVP binding behaviour during vegetative and reproductive stages

During the vegetative stage SVP acts as a repressor of the floral transition [6,8,18,19], while later it plays an important role during floral meristem specification and organogenesis by regulating expression of organ identity genes [20,25,48]. Here SVP binding sites were identified in seedlings and inflorescences to compare its behaviour at these two stages. A small number of direct target genes of SVP were previously identified in both vegetative and reproductive tissues [6,18,24]. Binding of SVP to these known sites was confirmed in the ChIP-seq data in both conditions, although in some cases (for example *SOC1* in vegetative tissues, see below) the enrichment after the IP was not sufficient to exceed the *P* value threshold employed.

The high confidence lists of putative targets of SVP in vegetative and reproductive tissues show a significant overlap, even if this does not imply a perfect overlap of binding regions for common target genes, as shown in the next section. In total 689 genes appear in both lists, which represents a highly significant overlap (*P* value < 1E-200) (Figure 3a and Additional data file 2, Table S2). The GO analyses reveal that the biological processes enriched in both stages are related to development, cell cycle, and DNA metabolism. These may define a set of genes that reflect the core role of SVP during plant development (Figure 3b).

SVP directly binds to flowering-time genes of different regulatory pathways

Mutations in *SVP* cause early flowering, illustrating a role for SVP in repressing the floral transition, a process controlled by several regulatory pathways [6,8]. Consistent with this function, GO terms related to development, such as ‘reproduction’ and ‘flower development’, are significantly overrepresented in the list of putative SVP targets (Figure 3). Moreover, SVP represses flowering by reducing the mRNA levels of *FT* and *TSF* [6,19] key components of the photoperiodic pathway, and of the floral integrator *SOC1* [18]. In the ChIP-seq data, *FT* is indeed bound by SVP, but with a low *P* value (9.5×10^{-7}) (data not shown). Similarly, ChIP-chip experiments performed by Tao and collaborators were not sensitive enough to detect the binding of SVP to the *FT* locus [37]. Recent

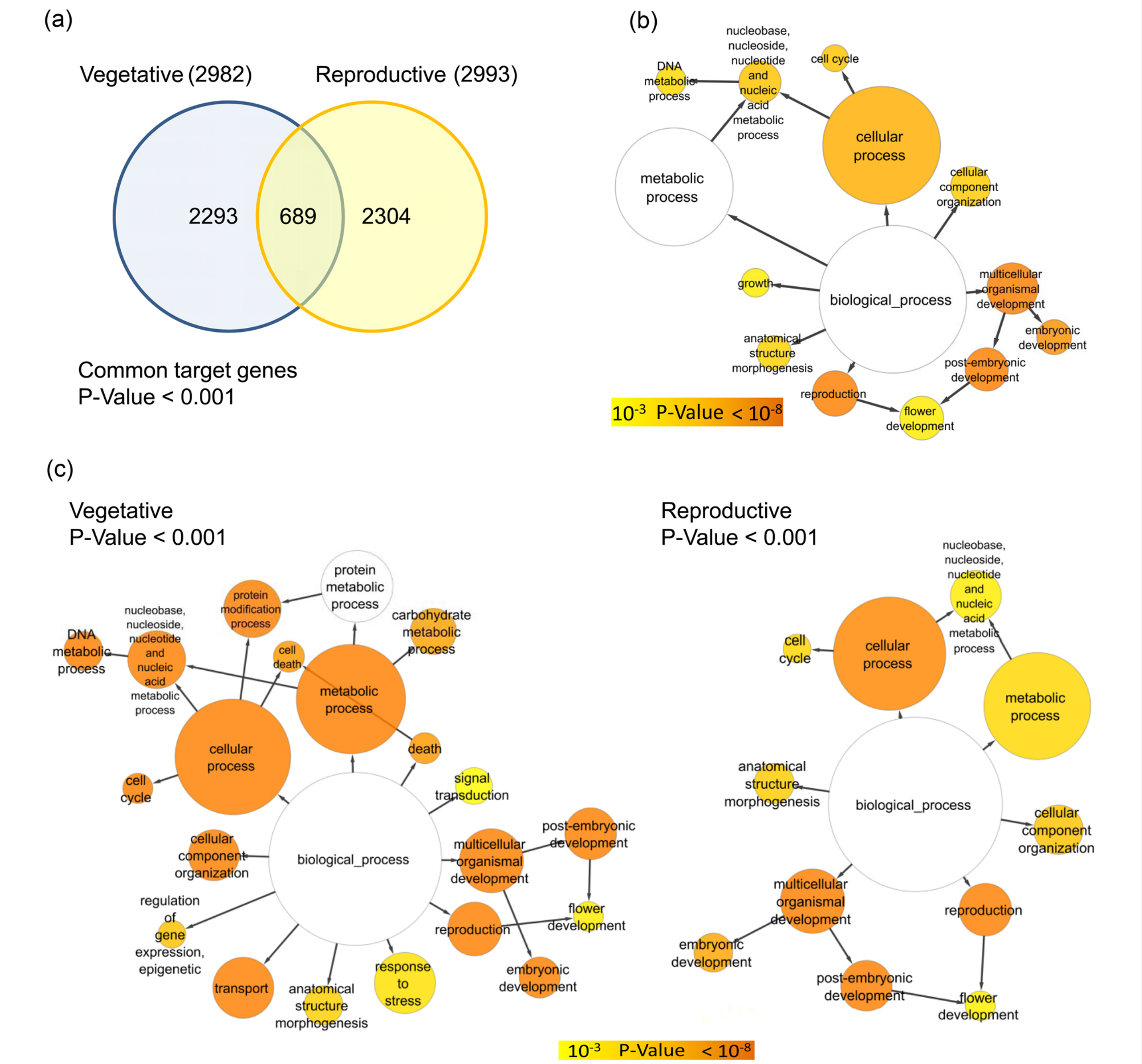


Figure 3 Common targets of SVP in vegetative and reproductive tissues and GO enrichment analysis of targets of SVP in vegetative and reproductive tissues. (**a**) Overlap between high confidence targets of SVP in vegetative and reproductive tissues; (**b**) GO enrichment analysis of biological process for the common targets of SVP in both tissues. Significant enrichments in the dataset were highlighted in color, different colors represented different levels of significance. The FDR was set up to 0.001 and 0.05; (**c**) visualization of the GO terms related to biological processes enriched in the high confidence lists of targets of SVP in vegetative (left) and reproductive (right) tissues. The color scale represents the level of significance, in terms of FDR, for each category.

work demonstrated that SVP also regulates flowering time independently of *FT* and *SOC1* [18,19]. Thus, we searched the list for known flowering-time regulators. Surprisingly, SVP bound genes involved in several different pathways (Additional data file 1, Table S3), including the circadian clock and photoperiodic pathway, represented by *GIGANTEA* (*GI*) and *PSEUDO-RESPONSE REGULATOR 7* (*PRR7*), the autonomous pathway, represented by genes such as *FLOWERING LATE KH MOTIF* (*FLK*) and *FLOWERING LOCUS D* (*FLD*), genes encoding components of chromatin associated complexes, such as *CURLY LEAF* (*CLF*), *SWINGER* (*SWN*), and *VERNALIZATION2* (*VNR2*), and the light signaling pathway represented by *PHYTOCHROME A* (*PHYA*).

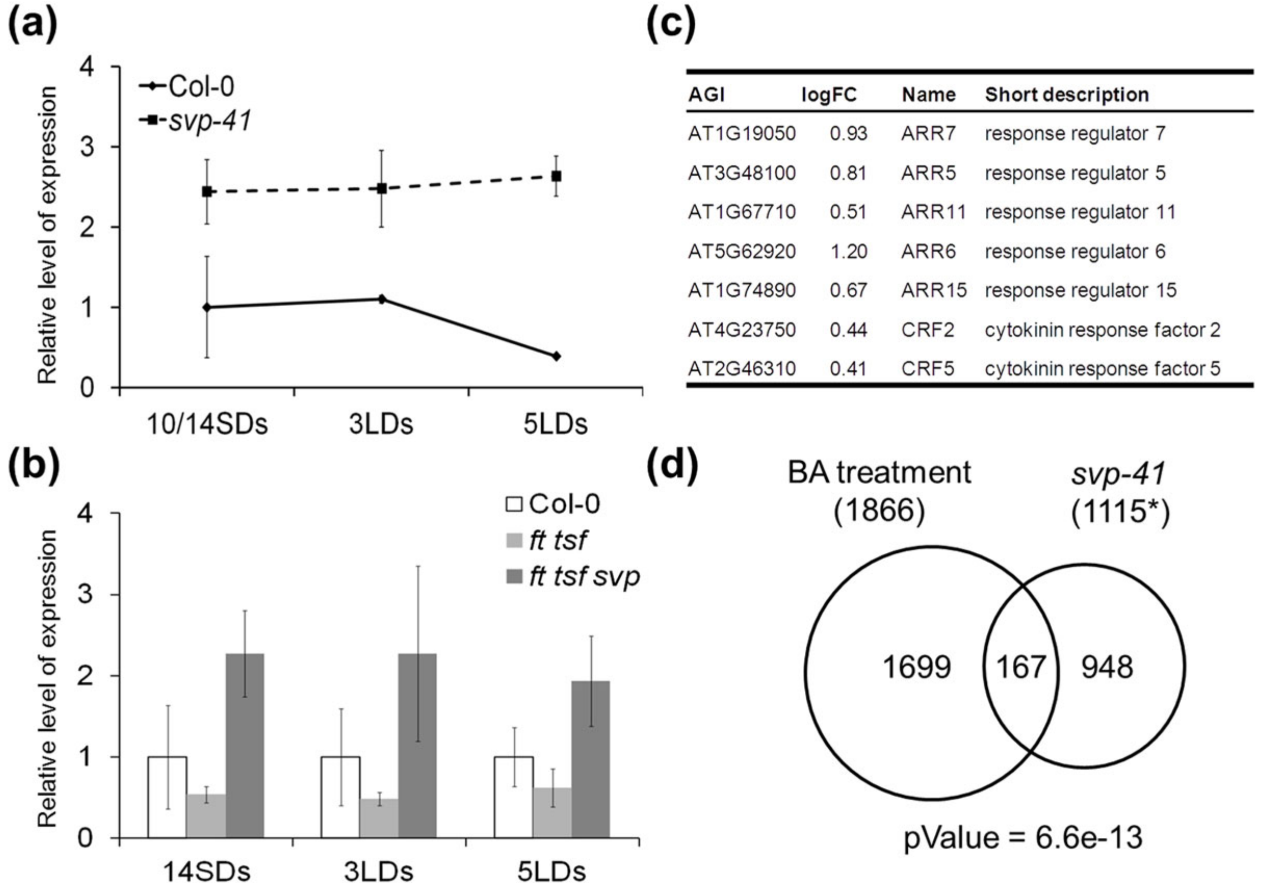
SVP and the regulation of growth regulator signaling during vegetative development

Growth regulators play different roles in flowering-time control and their molecular links to floral homeotic genes have been extensively reported [13,31,32]. SVP targets related to growth regulator signaling, response, transport and metabolism were identified in the ChIP-seq data (Additional data file 3, Table S4). For example, SVP binds directly to *STIP* (*STIMPY*), which was recently described as a component of the cytokinin (CK) signaling pathway [49], during the vegetative phase. The expression levels of this gene were tested in *svp-41* mutants and Col-0. The qRT-PCR experiments showed that *STIP* mRNA was present at significantly higher levels in *svp-41* mutants compared to Col-0 at all time points tested (Figure 4a). We also quantified the expression levels of *STIP* mRNA in *ft-10 tsf-1 svp-41*, which harbours null alleles of *FT* and *TSF* and *SVP* [19]. In *ft-10 tsf-1 svp-41* the expression levels of *STIP* were up-regulated compared to *ft-10 tsf-1* double mutants and Col-0 wild-type (Figure 4b), indicating that SVP controls this gene independently of the FT TSF photoperiodic signals. The effect of SVP on *STIP* expression might indirectly influence the expression of other genes involved in cytokinin signaling. To investigate this possibility a transcriptome analysis was performed by hybridizing RNA extracted from seedlings of wild-type Col-0 and *svp-41* to Affymetrix tiling arrays. The results of these experiments demonstrated that 1,381 genes were differentially expressed (FDR ≤0.05) in *svp-41* compared to Col-0 seedlings (Additional data file 4, Table S5). For some of these genes the change in expression in *svp-41* compared to Col-0 was also confirmed by qRT-PCR (Additional data file 1, Figure S2). A GO term test indicated that there is a significant enrichment of genes included in the category 'response to hormonal stimuli' (Additional data file 1, Figure S3 and Table S6). Interestingly seven genes upregulated in *svp-41* mutant were related to cytokinin signaling (Figure 4c). These genes belong to two different groups of cytokinin response genes: the type-A *ARABIDOPSIS RESPONSE REGULATORS* (*ARRs*) and the *CYTOKININ RESPONSE FACTORS* (*CRFs*). These two groups of genes are also transcriptionally activated by *STIP* [49], suggesting that the control of *STIP* by *SVP* has a broad effect on the cytokinin signaling pathway. Indeed, the effect of SVP on CK signaling was also reflected by the significant overlap (*P* value = 6.6×10^{-13}) between the lists of differentially expressed genes in *svp-41* mutant and the available expression-profiling data of seedlings treated with the CK benzyladenine (BA) [50] (Figure 4d and Additional data file 5, Table S7).

The ChIP-seq and tiling array data also suggested links between SVP and other growth regulators. For instance, SVP bound several genes involved in auxin signal transduction, such as *BIG*, which encodes a putative auxin transporter required for normal auxin efflux and inflorescence development (Additional data file 3, Table S4) [51,52]. Another gene bound by SVP is *CORONATINE INSENSITIVE 1* (*COI1*), which encodes the jasmonate receptor (Additional data file 3, Table S4) [53,54]. Therefore SVP might affect auxin and jamonate homeostasis by directly binding to genes encoding key components of their signaling cascade pathways. In agreement with this conclusion, our Tiling array data showed that members of the *SAUR-like auxin-responsive* family were up-regulated in *svp-41* mutant (Additional data file 3, Table S4 and Additional data file 1, Figure S2). In addition, six of the *JASMONATE ZIM-domain* (*JAZ*) genes (*JAZ1*, *5*, *6*, *7*, *8*, and *10*), which are part of the jasmonate signaling pathway and are transcriptionally activated by the hormone, were increased in expression in the mutant compared to *Col-0* (Additional data file 3, Table S4 and Additional data file 1, Figure S2).

Common targets of SVP and FLC during vegetative development

MADS-domain proteins form multimeric complexes that are proposed to be important in determining their DNA binding specificity. Co-immunoprecipitation analysis and yeast two-hybrid assays demonstrated that SVP interacts with the related MADS-domain protein FLC and genetic data indicate that this interaction is likely functionally important in the control of flowering time [18,55]. Moreover, SVP associates with the promoter region of *SOC1* and the intron of *FT* where FLC also binds [18,39]. Recently the genome wide targets of FLC were identified using ChIP-seq technology [32]. Of these FLC putative targets, 112 were also detected in our experiment as being bound by SVP in vegetative tissue (*P* value = 1.9×10^{-6}) (Additional data file 1, Figure S4a). Nine of the FLC putative targets were previously validated by ChIP-qPCR and six of them shown to change in expression in *flc-3* mutants [32]. Of these confirmed FLC targets, four were selected to test by ChIP-qPCR if they were also bound by SVP (Figure 5b, c). Of these four FLC targets, three were bound by SVP in a similar location. One of these was *JAZ6*, which was bound by FLC in its promoter region and its expression is increased in *flc-3* [18]. *JAZ6* expression was also upregulated in *svp-41* (Figure 5a), however it was not enriched in our ChIP-seq experiment, and this was confirmed by independent ChIP-qPCR analysis, suggesting that the changes in *JAZ6* expression caused by SVP are not an effect of direct binding (Figure 5c). A second confirmed FLC target, *AGL16*, was not enriched in the SVP ChIP-seq data, however the region bound by FLC showed a low but consistent enrichment in ChIP-qPCR of SVP. This experiment suggests that SVP is weakly bound to the same region of *AGL16* as FLC, and



AGI	logFC	Name	Short description
AT1G19050	0.93	ARR7	response regulator 7
AT3G48100	0.81	ARR5	response regulator 5
AT1G67710	0.51	ARR11	response regulator 11
AT5G62920	1.20	ARR6	response regulator 6
AT1G74890	0.67	ARR15	response regulator 15
AT4G23750	0.44	CRF2	cytokinin response factor 2
AT2G46310	0.41	CRF5	cytokinin response factor 5

Figure 4 SVP regulates *STIP* and genes involved in cytokinin signaling. (**a**) Quantification by qRT-PCR of the *STIP* mRNA levels in *svp-41* (dashed line) compared to Col-0 (solid line). *svp-41* and Col-0 seedlings were grown for 10 and 14 days, respectively, under SDs and then shifted to LDs for 3 or 5 days; (**b**) expression levels of *STIP* in Col-0, *ft-10 tsf-1*, and *ft-10 tsf-1 svp-1* genetic backgrounds. Seedlings were grown under SDs for 14 days and then shifted to LDs for 3 or 5 days. Bars in (a) and (b) show standard deviation of the mean of three biological repeats; (**c**) list of genes differentially expressed in *svp-41* compare to Col-0 related to the CK signaling pathway; (**d**) comparison between genes affected by BA treatment (Brenner *et al*., 2005) and genes de-regulated in *svp-41*. For this analysis the genes represented in the affymetrix tiling Array 1.0R (used in the present study), but not represented in the ATH1 microarray used by Brenner *et al*. (2005), were excluded. A total of 167 genes were common between the two conditions ($P = 6.6 \times 10^{-13}$, based on the binomial test).

the low enrichment might explain why it was not detected in the ChIP-seq experiment. *AGL16* expression was not changed in *svp-41* compared to Col, similar to what was observed in *flc-3*. A third confirmed FLC target was SVP, and ChIP-qPCR confirmed that SVP binds to the same region in its own promoter as FLC. These ChIP-qPCR experiments demonstrate that there is a strong but not complete overlap in the targets of FLC and SVP.

SVP auto-regulates its gene expression in vegetative tissue and flowers

The ChIP-seq data indicated that SVP binds to its own genomic region in vegetative tissue and flowers. However, regions actually bound in both tissues may differ. This differential binding was confirmed by independent ChIP-qPCR experiments on two specific regions named I and II (Figure 6 a-c), located approximately 2,000 bp upstream of the 5'UTR and in the terminal part of the *SVP* first intron, respectively. As shown in Figure 6b and 6c, SVP binds site I in floral tissue but not in vegetative tissue, whereas site II is bound in both tissues. Whether binding of SVP influenced its own expression was tested in different ways. In addition to the microarray experiment described above, another transcriptome analysis was performed by hybridizing RNA extracted from inflorescences of wild type Col-0 and *svp-41 agl24 ap1-12* to affymetrix tiling arrays. In this experiment 246 genes were differentially expressed (FDR ≤0.05) in *svp-41 agl24 ap1-12* compared to Col-0 inflorescences (Additional data file 4, Table S5). The tiling array expression data showed that *SVP* mRNA was downregulated in the *svp-41* single mutant in vegetative tissues (logFC -1.13; *P*=0.001) as well as in inflorescences of the *svp-41 agl24-2 ap1-12* triple mutant (logFC -0.86;

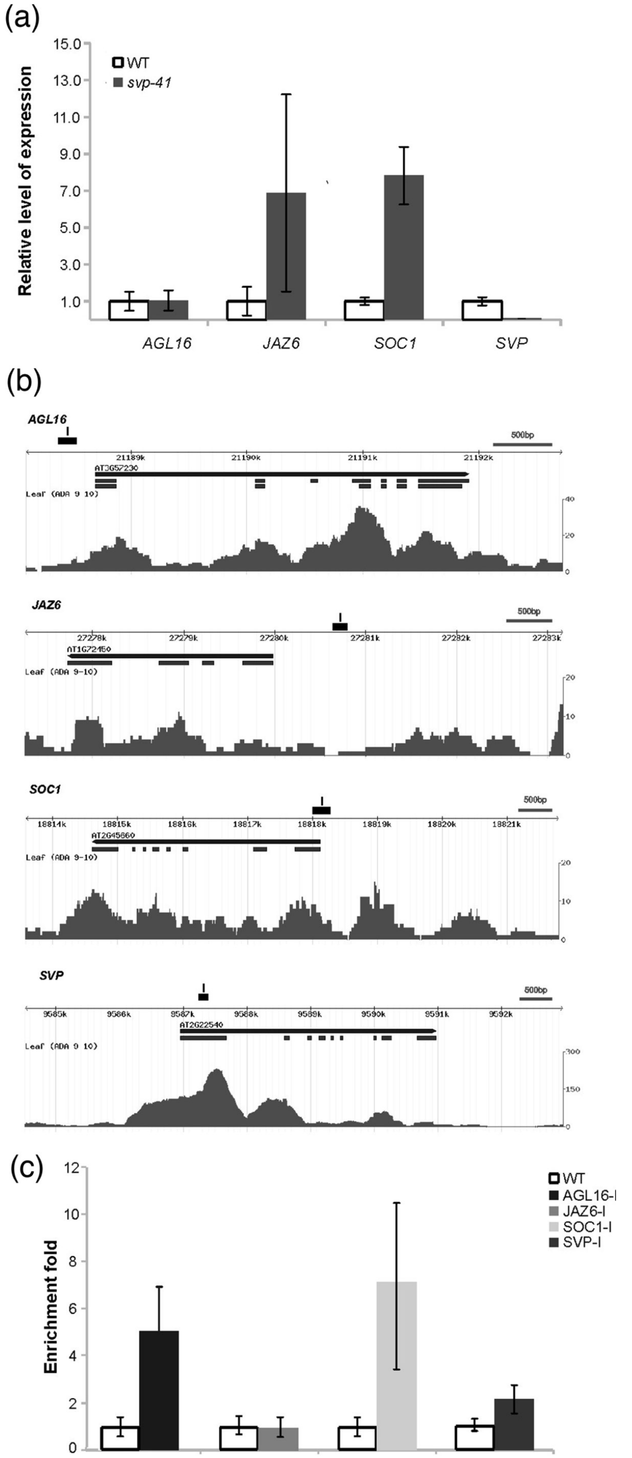


Figure 5 Common targets of SVP and FLC. (**a**) Expression of known direct targets of FLC in *svp-41*. Data represent expression of selected genes in microarray experiment with FDR <0.05. The expression level of each gene in *svp-41* was normalized to the level of wild type Col-0. Error bars represent SDs of normalized data; (**b**) binding profiles of ChIP-seq experiment for the selected genes. TAIR annotation corresponds to TAIR8. Grey boxes represent the region validated by ChIP-PCR which are shown in panel (**c**); (c) ChIP-PCR validation of selected genes using anti-GFP antibodies using seedlings of wild type Col-0 and SVP::SVP-GFP *svp-41* lines. Results are expressed relative to actin. Error bars represent SD.

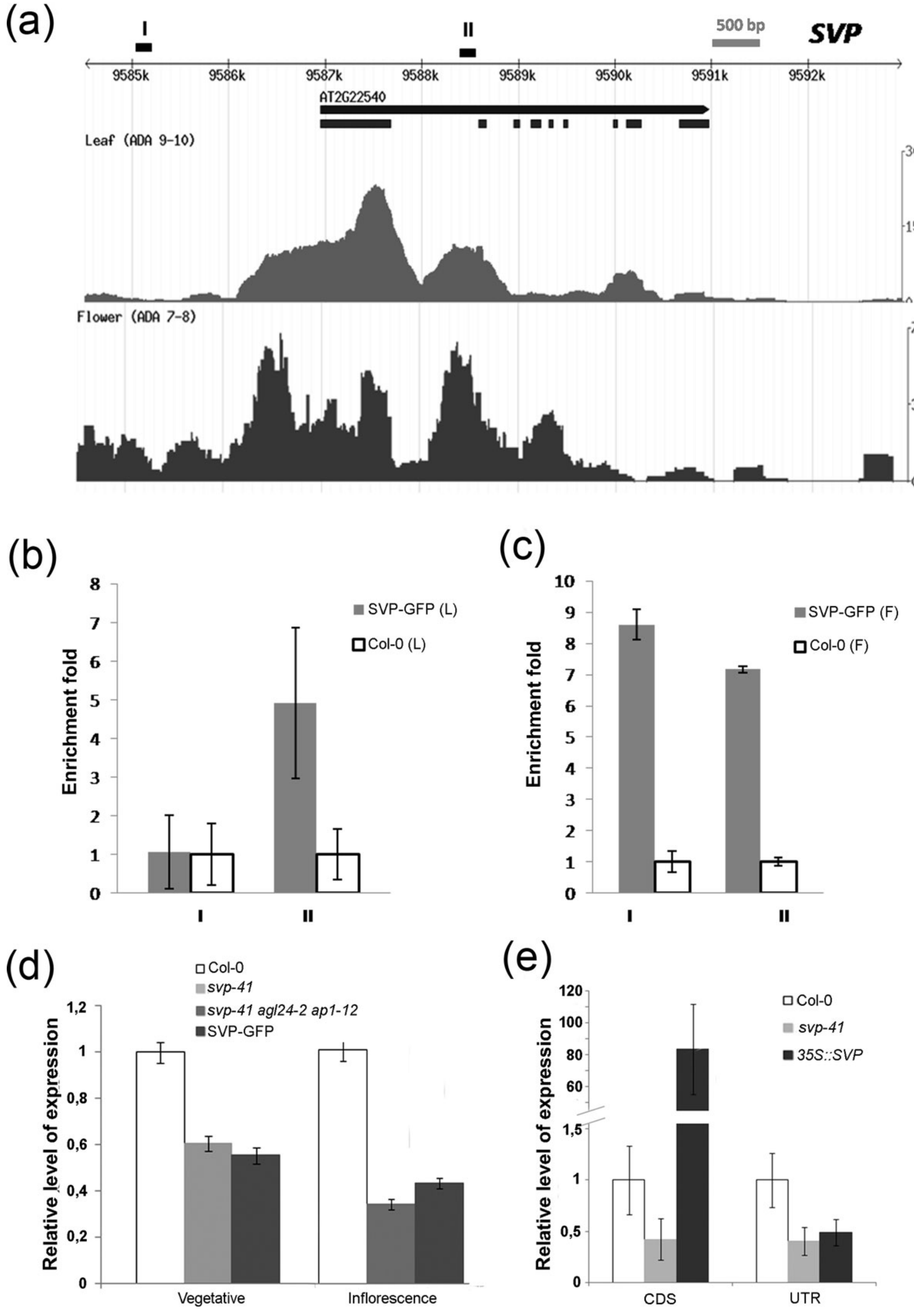


Figure 6 SVP binds and regulates itself. (**a**) Binding profiles for SVP on *SVP* genomic locus in seedlings (upper panel) and inflorescence (lower panel) tissues. TAIR annotation corresponds to TAIR8. Grey boxes represent the region validated by ChIP-PCR in panels (**b**) and (**c**); (b) and (c) ChIP-PCR validations for two specific regions named I and II. ChIP assays were done using GFP antibodies and SVP::SVP-GFP *svp-41* plants and compared to wild-type control plants. ChIP-PCR validation in vegetative (b) and reproductive tissue (c); (**d**) qRT-PCR expression analysis using primers for the *SVP* 3'UTR region. RNA was extracted from wild-type Col-0, *svp-41*, and SVP::SVP-GFP *svp-41* seedlings and from wild-type Col-0, *svp-41 agl24-2 ap1-12* triple mutant, and SVP::SVP-GFP *svp-41* inflorescences; (**e**) qRT-PCR expression analyses using primers for the *SVP* 3'UTR region and coding region. RNA was extracted from wild-type Col-0, *svp-41*, and 35S::*SVP* seedlings. In all graphs error bars represent the standard deviation of normalized data (SD).

P=0.02). This downregulation was validated by qRT-PCR using independent *svp-41* single mutant, *svp-41 agl24-2 ap1-12* triple mutant and wild-type cDNA samples obtained from RNA extracted from seedlings and inflorescences (Figure 6d). Since *svp-41* is a deletion mutant in which two base pairs are deleted in the second exon resulting in a frame-shift of the open reading frame [8], this reduction in mRNA level might be due to nonsense-mediated decay [56]. To investigate this possibility, we performed qRT-PCR assays using primers designed on the 3'UTR region of the endogenous *SVP* gene, which is not present in the *SVP::SVP-GFP* fusion construct. RNA was extracted from wild-type, *svp-41* and *SVP::SVP-GFP svp-41* seedlings and from wild-type, *svp-41 agl24-2 ap1-12* and *SVP::SVP-GFP svp-41* inflorescences (Figure 6d). The results confirmed a reduction in mRNA level also in *SVP::SVP-GFP svp-41* tissues suggesting that indeed this reduction in *SVP* mRNA level seems to depend on the mRNA instability in the mutant background. As an alternative approach the abundance of *SVP* mRNA expressed from the endogenous gene was tested in plants in which *SVP* was overexpressed from a *35S::SVP* transgene. A qRT-PCR strategy was used in which the cDNA expressed from the transgene and endogenous gene can be distinguished (Figure 6e). This experiment demonstrated that *SVP* mRNA expressed from the endogenous locus is reduced in *35S::SVP* plants. Taken together our data suggest that SVP directly regulates its own expression, and that it probably acts to repress its own transcription.

Genes involved in meristem development are targets of SVP at two developmental stages

Genes involved in meristem development were enriched as SVP targets in both vegetative material and flowers. SVP is expressed in the SAM during the vegetative stage [6,8,18,19]. In addition it plays an important role during floral meristem specification and organogenesis [25,48]. Consistent with this idea a significant enrichment of SVP target genes related to post-embryonic developmental processes was detected in the ChIP-seq results of both vegetative and reproductive samples (Figure 3c). Due to the expression pattern of *SVP*, putative targets with annotated functions in meristem development were screened for directly (Additional data file 3, Table S4). The *CLV-WUS* feedback loop plays a central role in maintaining meristematic activities [57]. In the ChIP-seq data *CLV1* and *CLV2*, two important players in *WUS* regulation, are targets of SVP in vegetative tissues and *CLV1* is also bound during reproductive development. Additionally, according to the ChIP-seq data, the HD-ZIPIII encoding genes *PHABULOSA* (*PHB*), *PHAVOLUTA* (*PHV*), *REVOLUTA* (*REV*), and *HOMEOBOX GENE 8* (*ATHB8*), which regulate post-embryonic meristem initiation [58], are also bound by SVP in vegetative tissue. Furthermore, *PHB* which is a regulator of the size of the *WUS*-expression domain [59], is also bound by SVP in the floral meristem. In order to test whether the binding of SVP to some of these genes affects their spatial pattern of expression we performed RNA *in-situ* hybridization experiments. A broader expression pattern of *PHB* and *CLV1* was observed in shoot apical meristems of *svp-41* mutants than Col-0 wild-type plants grown for 2 weeks under SDs (vegetative phase) (Figure 7a, b, d, e). However, these differences might be due to the larger size of the *svp-41* meristem compared to Col-0 at this stage. Thus, the patterns of expression of *PHB* and *CLV1* were also compared in 10-day-old *svp-41* mutants and 2-week-old Col-0 plants, which have SAMs of similar size. Confirming our previous result *PHB* and *CLV1* mRNA were detected in a broader region of the *svp-41* (10 SDs) SAM compared to Col-0 (Figure 7c and 7f). These results together with the ChIP-seq data suggest that SVP directly regulates the expression pattern of these genes. Furthermore, *KANADI1* (*KAN1*) and *KAN2*, involved in the establishment of abaxial-adaxial polarity in lateral organs produced from the apical meristem, resulted also to be direct targets of SVP in inflorescences. It has been hypothesized that complementary regions of action of the class III HD-ZIP genes and *KANADI* genes leads to the establishment of adaxial and abaxial domains in developing lateral organs. The possible role of SVP and other MADS-domain proteins in the regulation of part of these genes in reproductive tissues is presented below.

Genome wide targets of SVP during flower development and comparison with the targets of AP1 and SEP3

During the early stages of flower development (stage 1 and 2) AP1 interacts with SVP and the dimer recruits the SEU-LUG repressor complex to control the expression of homeotic genes to maintain the floral meristem in an undifferentiated state [25]. At late stage 2, when *SVP* expression is switched off, AP1 interacts with SEP3 to control sepal and petal identity. Recently, genome-wide binding studies for SEP3 and AP1 during inflorescence development were published [13,31] providing the opportunity to compare these datasets with the one obtained here for SVP.

A total of 265 common putative targets for both SVP and AP1 were identified (*P* value <7.2E-06) (Additional data file 6, Table S8 and Additional data file 1, Figure S4). This overlap is expected because SVP and AP1 act redundantly during floral meristem specification where their expression domains overlap [24]. Interestingly transcription factors are enriched among common targets. In addition SVP binds to *AP1*, suggesting that it regulates a

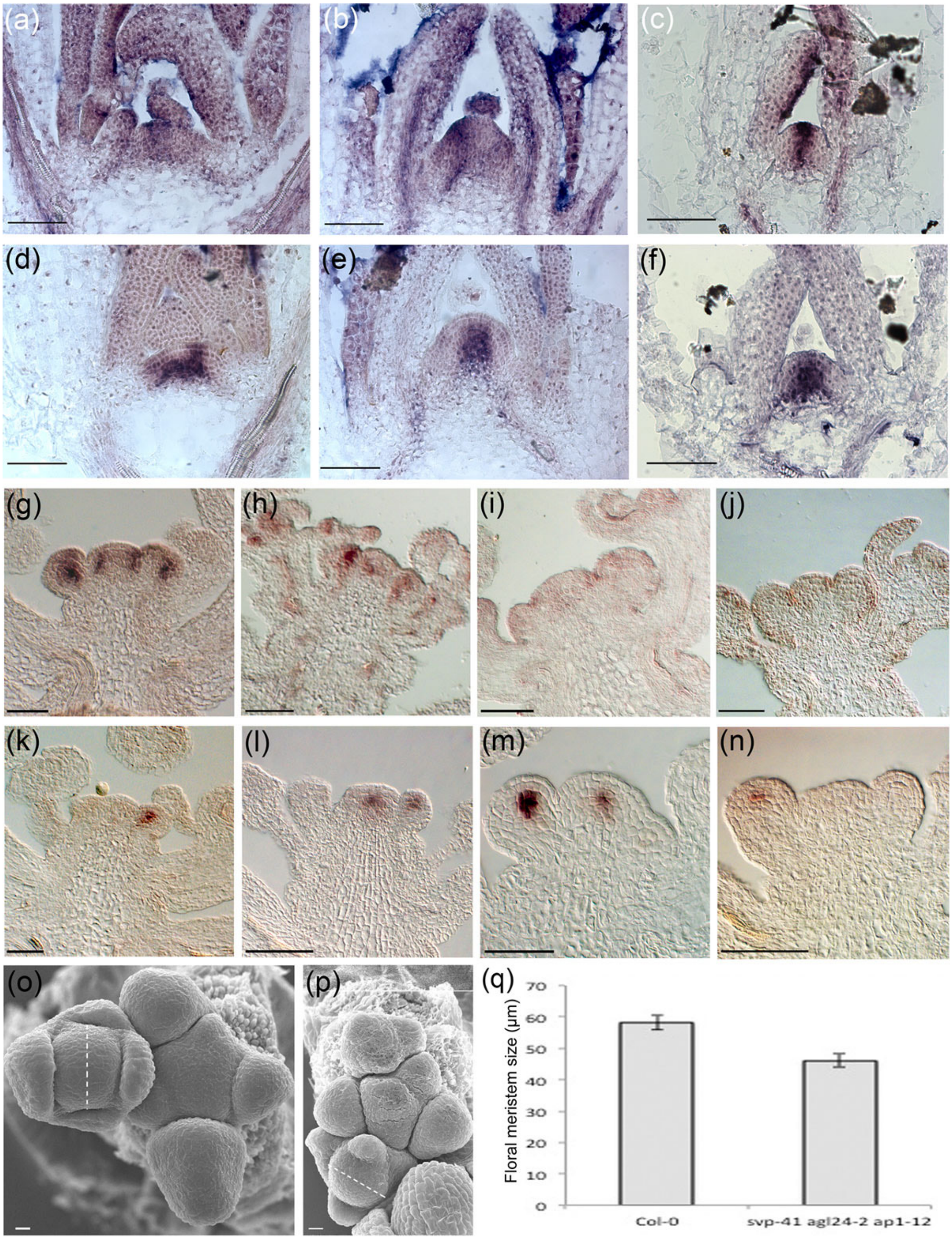


Figure 7 Expression analysis of meristem developmental genes by *in-situ* hybridization analysis in vegetative and reproductive tissues and floral meristem size analysis. (**a-c**) Patterns of expression of *PHB*: (a) 14-day-old wild-type, (**b**) 14-day-old *svp-41*, and (c) 10-day-old *svp-41* mutant; (**d-f**) patterns of expression of *CLV1:* (d) 14-day-old wild-type, (**e**) 14-day-old *svp-41* mutant, and (f) 10-day-old *svp-41* mutant; in both *svp-41* 10 and 14-day-old seedlings the *PHB* and *CLV1* mRNA were detected in a broader region of the SAM compared to Col-0; (**g, h**) expression of *ARF3* in wild type and *svp-41 agl24-2 ap1-12* inflorescence respectively; (**i, j**) *KAN1* expression pattern in wild-type and *svp-41 agl24-2 ap1-12* inflorescences; (**k, l**) *CLV1* expression in wild-type and *svp-41 agl24-2 ap1-12* inflorescence; (**m, n**) expression profile of *WUS* in wild-type and *svp-41 agl24-2 ap1-12* inflorescences, its expression seems to be higher in the wild-type FM than in the triple mutant FMs at the same developmental stage. The scale bar represents 50 μm. (**o**) View of wild-type inflorescence; (**p**) view of *svp-41 agl24-2 ap1-12* inflorescences; central zone of triple mutant FMs at stage 3 were compared to those of wild-type plants. The scale bar represents 10 μm. (**q**) Diagram showing the difference in FMs size between the wild-type and *svp-41 agl24-2 ap1-12* triple mutant central dome, error bars represent standard error (SE).

functionally redundant gene as well as itself. The overlap between the targets of SVP with those published for SEP3 [13] revealed 413 (*P* value <5.91E-10) genes that are bound by both of these MADS domain transcription factors (Additional data file 6, Table S8 and Additional data file 1, Figure S4). *KAN1*, *CLV1*, *PHB*, and *ARF3* also named *ETTIN*, that are present in the subset of genes bound by SVP and AP1, are also present in the list of genes regulated by both SVP and SEP3.

Transcriptome data obtained from the tiling array hybridization experiments using RNA extracted from inflorescences of Col-0 and the *svp-41 agl24 ap1-12* mutant showed that the number of differentially expressed genes were considerably fewer than those found by comparing the vegetative tissue of *svp-41* and Col-0 wild-type plants (Additional data file 4, Table S5). However, the number of deregulated genes might be underestimated in this analysis because the whole inflorescence of *svp-41 agl24 ap1-12* mutant plants were used, whereas *SVP* expression is restricted to stage 1-2 FMs only. Therefore, altered expression of several targets might not be detected in this material. To overcome this we also checked the expression of putative SVP target genes by a qRT-PCR approach, collecting the most inner parts of *Col-0* and *svp-41 agl24 ap1-12* inflorescences, avoiding the already opened flowers. Both *KAN1* and *PHB* mRNAs were increased in abundance in the *svp-41 agl24-2 ap1-12* mutant background compared to wild-type (Figure 8a) and the enrichment of these genes observed in the ChIP-seq experiment was confirmed by means of independent ChIP-qPCR analysis (Figure 8b and 8c), suggesting a direct regulation of these genes by both SVP and AP1 during flower development. Interestingly also *CLV1*, which plays an important role in establishing and maintaining floral meristem identity [60], is a direct target of both SVP and AP1 in reproductive tissue and its expression was increased in the *svp-41 agl24-2 ap1-12* triple mutant compared to wild-type (Figure 8a-c). Another transcription factor encoding gene that is bound by SVP and AP1 and upregulated in *svp-41 agl24-2 ap1-12* inflorescences is *ARF3* (Figure 8a-c). ARFs are proteins that are activated by convergent auxin flow. Dynamic changes in auxin fluxes are mediated by PIN proteins and interestingly SVP and AP1 can interact with the genomic region of *PIN1*. Analysis by qRT-PCR showed increased levels of *PIN1* mRNA in *svp-41 agl24-2 ap1-12* inflorescences in comparison to the wild-type control, suggesting a direct role of SVP and AP1 in its regulation which was confirmed by independent ChIP-qPCR experiments (Figure 8a-c). We further examined the expression of *ARF3*, *CLV1*, *KAN1*, *PHB*, and *PIN1* in response to SVP activation using the functional steroid-inducible system. The *svp-41 agl24-2 ap1-10* triple mutant was transformed with a construct in which the 35S promoter directs a fusion between SVP and a part of the rat glucocorticoid receptor (GR), as reported previously [61]. The *svp-41 agl24-2 ap1-10* mutant forms cauliflower like curds since its unable to establish FM identity and therefore it proliferates IMs instead. The obtained transgenic plants showed upon induction with the steroid dexamethasone (DEX) rescue of the development of FMs and flowers that resembled those of the *agl24-2 ap1-10* double mutant (Additional data file 1, Figure S5). We treated the inflorescences twice, at time 0 and again after 8 h with DEX and collected the material after 24 h from the first treatment. This time point was selected according to Smyth *et al.* [40], since they showed that the duration of stage 1 of flower development is 24 h. *ARF3*, *CLV1*, *KAN1*, *PHB*, and *PIN1* expression levels were all decreased after DEX treatment of *svp-41 agl24-2 ap1-10* 35S::SVP-GR inflorescences, confirming that SVP acts as a repressor of those genes (Figure 8d).

To investigate the changes in expression profiles of some of these target genes, we performed *in-situ* hybridization experiments using wild type and *svp-41 agl24-2 ap1-12* inflorescences (Figure 7g-n). For *ARF3*, *KAN1*, and *CLV1* the expression pattern was not changed suggesting that the upregulation of these genes is not due to ectopic expression. Interestingly *in situs* using a specific probe for *WUS* clearly showed that in comparison to wild-type, in stage 2 FMs this gene was lower expressed in the *svp-41 agl24-2 ap1-12* triple mutant. Since *svp-41 agl24-2 ap1-12* flowers show reduced numbers of floral organs compared to wild-type or any of the single mutants [25], we wondered if these defects were caused by changes in meristem size. Therefore the central zone of FMs at stage 3 of flower development of the *svp-41 agl24-2 ap1-12* triple mutant and wild-type were compared. The size of the central zone is defined by the distance between the opposite lateral sepals (Figure 7o-q). The *svp-41 agl24-2 ap1-12* FMs were significantly smaller, as compared to those of wild-type plants (Table 1 and Figure 7q). Taken together all these data suggest a role of SVP in the control of FM size, probably by modulating the expression of genes involved in the CLV-WUS pathway.

SVP binds in reproductive tissues to genes encoding post-translational regulators

Interestingly, the high confidence list of SVP target genes in inflorescence tissue exhibits a significant enrichment of genes related to Cullin-RING ubiquitin ligase complexes, mainly involved in post-translational regulation of substrate proteins by attaching poly-ubiquitin chains that target the substrate for 26S proteasome degradation [62,63]. The substrate specificity of CUL4-RING-LIGASES (CRL4s) is exerted by proteins that contain a DWD box (DDB1-binding WD-40 box)

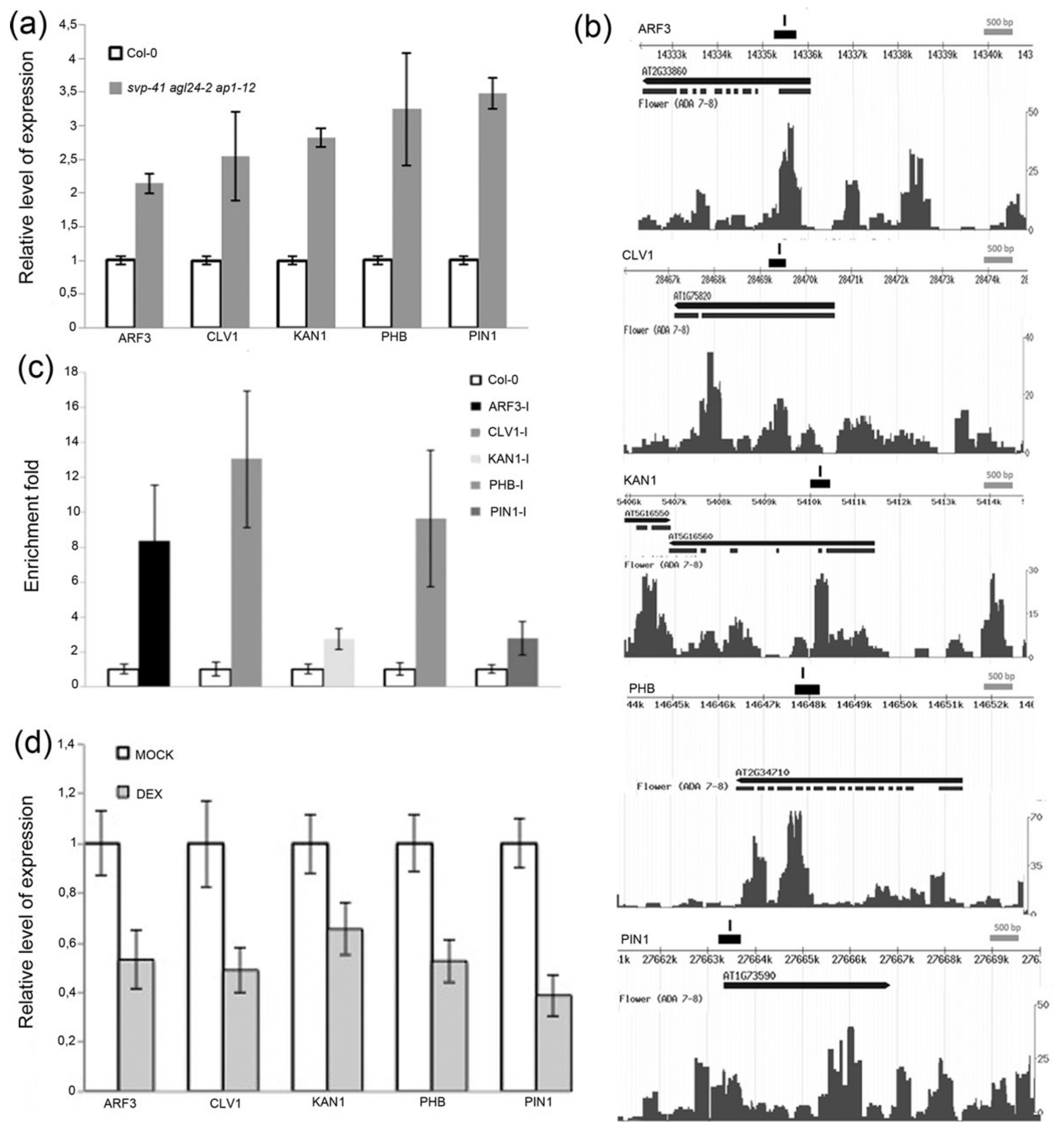


Figure 8 Common targets of SVP and AP1. (**a**) Expression analyses of *ARF3*, *CLV1*, *KAN1*, *PHB*, and *PIN1*. RNA was extracted from wild-type Col-0 and *svp-41 agl24-2 ap1-12* triple mutant inflorescences. (**b**) Binding profiles of ChIP-seq experiment using inflorescence tissue for the selected genes. TAIR annotation corresponds to TAIR8. Black boxes represent the region validated by ChIP-PCR, which are shown in (**c**). (c) ChIP-PCR validation of selected genes using anti-GFP antibodies and inflorescences of wild-type Col-0 and the SVP::SVP-GFP *svp-41* line. Error bars represent standard deviations (SD) of normalized data; (**d**) relative level of expression of *ARF3*, *CLV1*, *KAN1*, *PHB*, and *PIN1* in *svp-41 agl24-2 ap1-10, 35S: SVP-GR* plants that were mock-treated or with 10 μM dexamethasone.

or a WDxR sub-motif [64-67]. Proteins with these motifs are referred to as potential DCAF (DDB1-CUL4 ASSOCIATED FACTOR) proteins [67], which may target proteins for ubiquitinilation [64,68]. However, they have also been implicated in chromatin mediated transcriptional control [69]. In Arabidopsis, 119 different putative DCAF proteins have been identified [67] and our ChIP-seq experiments suggest that nearly half of

Table 1 Floral meristem size

	Floral meristem mean ±SE (μm)
Col-0	*n*=8 58.1 ± 2.2
svp-41 agl24-2 ap1-12	*n*=8 46.7 ± 2

Col-0 vs. svp-41 agl24-2 ap1-12: Two sample T-test, t = 3.9200, DF = 14, *P*=0.0015.

them (47 of 119) are targets of SVP in both tissues tested and more than half of these (26 of 47) are putative SVP targets in reproductive tissues (Additional data file 1, Table S9).

Among the putative DCAF floral SVP targets to which a function in floral development had not previously been ascribed (Additional data file 1, Table S9), we selected *WDR55* as a case study for detailed analyses of its function as a SVP target in flower development.

The regulation of *WDR55* by SVP forms as an alternative pathway for the regulation of *AG*

WDR55 was shown to interact with DDB1A, suggesting a regulative role through a putative CUL4-DDB1$^{\text{WDR55}}$ E3 complex, and plays a major role in *Arabidopsis* reproductive development. *WDR55* is required for gametogenesis and embryogenesis and is suggested to be involved in auxin-dependent regulation of embryo development [70].

In order to verify that *WDR55* expression requires *SVP*, we performed qRT-PCR analyses on *svp* double and triple mutant combinations. Compared to wild-type, *WDR55* transcripts were reduced in abundance in the double mutant *svp-41 agl24-2* (30°C) and in the *svp-41 agl24-2 ap1-12* mutant background (Figure 9a). The binding of SVP to *WDR55*, as observed in the ChIP-seq experiment, was confirmed by means of independent ChIP-qPCR analysis (Figure 9b), suggesting that changes of *WDR55* expression in *svp-41 agl24-2* and *svp-41 agl24-2 ap1-12* are due to the direct action of SVP during flower development.

A recent report describes two mutant alleles of *WDR55* that demonstrate a requirement of WDR55 in gametophyte development and function, as well as for setting up the embryo body plan. The weaker of these alleles, *wdr55-2*, displayed close to mendelian ratios of mutant seeds (22.7%) and no homozygous plants could be identified, although a small fraction (2%) could be expected from the genetic data [70]. In order to screen for the theoretical presence of homozygous plants in the progeny, we allowed a large number of seeds from heterozygous *wdr55-2* plants to germinate for a prolonged period on MS-2 agar plates containing glufosinate (BASTA) selection. Indeed, we identified a class of late germinating, small seedlings that initially were smaller than the glufosinate sensitive seedlings (3.6%, *n* = 1,035). However, this class was BASTA resistant and thus carried the *wdr55-2* mutation.

Generally, *wdr55-2* seedlings supported growth, but were severely delayed compared to wild-type. In particular, *wdr55-2* inflorescences were smaller than wild-type and had fewer flowers. Upon inspection we found that the mutant floral organs were generally smaller and often morphologically distinct from wild-type (Figure 9 and Additional data file 1, Figure S6). The sepals were thinner and often fused at early stages and did not separate completely at maturation (Figure 9d and 9e, Additional data file 1, Figure S6b, c and Table S10). The petals were smaller and thinner, as well as being non-uniform in size (Additional data file 1, Figure S6e and Table S10). The stamens were smaller and never occurred in sixes as in wild-type Col (Additional data file 1, Figure S6 and Table S10). The *wdr55-2* flowers also displayed homeotic transformations (Figure 9g, i). We observed unfused carpels (Figure 9f), carpeloid sepals (Figure 9g), petals that resemble stamens filaments and carpeloid filaments with ectopic papillar cells (Figure 9i) at a moderate frequency. New flowers appeared to grow out from whorl 1 or 2 at a low frequency (Figure 9j) and most of the flowers appeared to be asymmetric in flower organ organization (Figure 9h).

Due to the homeotic transformations observed in *wdr55-2* flowers, we checked the expression of the organ identity genes *APETALA3 (AP3)*, *PISTILLATA (PI)*, and *AGAMOUS (AG)* by *in-situ* hybridization (Figure 9k-n and Additional data file 1, Figure S7). The *in-situ* analysis shows that in the *wdr55-2* mutant, the expression pattern of both *AP3* and *PI* is maintained as wild-type plants (Additional data file 1, Figure S7).

AG is expressed in the inner part of the floral meristem where stamen and carpel primordia develop. During flower development *AG* expression is restricted to whorls 3 and 4 (Figure 9k). The *in-situ* analysis shows that in the *wdr55-2* mutant, *AG* is expressed in chimeric organs that develop in the second whorls (Figure 9l) as well as in carpelloid-sepals developing in first whorls (Figure 9m) where stigmatic tissues and carpelloid structures are detectable. *AG* is expressed already in early stages of flower development, in particular stage 1 (Figure 9n), but the architecture of inflorescences in *wdr55-2* makes precise staging difficult.

SEU, LUG, AP1, and SVP are involved in *AG* regulation, and by mutation ectopic *AG* expression is found [25,71-73]. *SEU* and *LUG* are thought to be cadastral genes, and are involved in the control of expression boundaries of floral homeotic genes [71,73] and they interact to repress *AGAMOUS* (*AG*) in the outer two whorls of the flower [72,73]. The SVP-AP1 dimer binds the LUG-SEU repressor and directly regulates *AG* expression during early stages of flower development

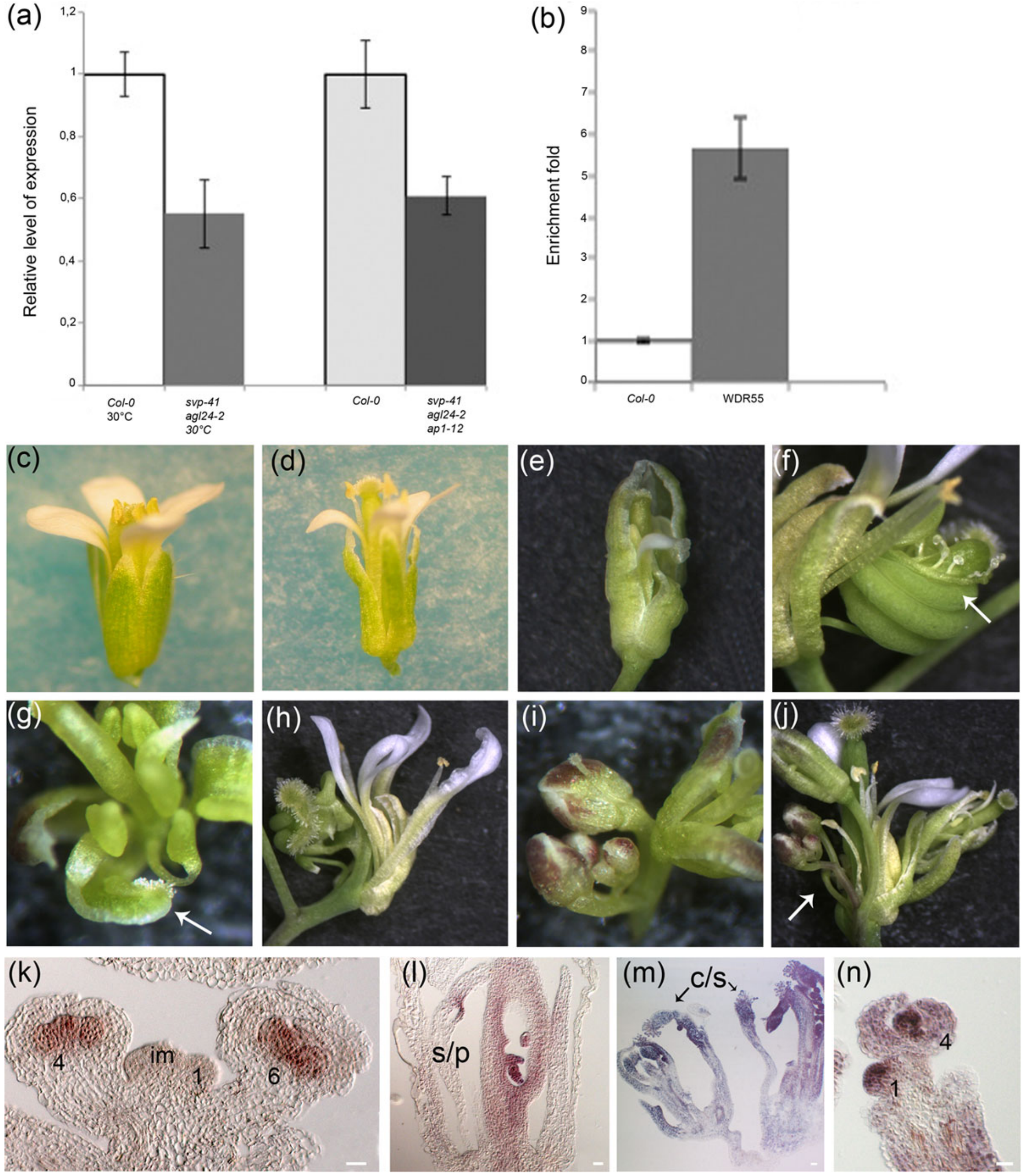


Figure 9 ***WDR55*** **is a target of SVP in reproductive tissues**. (**a**) Expression analyses of *WDR55*, RNA was extracted from wild type Col-0 and *svp-41 agl24-2 ap1-12* triple mutant inflorescences. (**b**) ChIP-PCR validation of *WDR55* region bound by SVP in reproductive tissues; (**c-j**) phenotypic analysis of (c) wild-type and (**d**-j) *wdr55-2* flowers: (d) flower of *wdr55-2* mutant showing reduced size of sepals and petals; (**e**) flower of *wdr55-2* mutant with unfused sepals, (**f**) arrow indicates unfused carpel, (**g**) arrow indicates chimeric sepal bearing stigmatic tissue on the top, (**h**) asymmetric development of *wdr55-2* flower, (**i**) *wdr55-2* flowers often develop new flowers in whorls 1 and 2; (**k-n**) *in-situ* hybridization of (k) wild-type, and (**l-n**) *wdr55-2* using *AG* specific probe: (k) in wild-type inflorescence *AG* is expressed from late stage 2 and its expression is restricted to the third and fourth whorl (im is inflorescence meristem), (l) ectopic expression of *AG* in staminoid petal of *wdr55-2* flower, (**m**) carpelloid sepals expressing *AG* in *wdr55-2* flower, (n) early stages of flower development in *wdr55-2* mutant in which the expression of *AG* is already detectable. The scale bar represents 20 μm.

[20,25]. To investigate the regulation of *AG* through WDR55 further, a Yeast-2-Hybrid (Y2H) was performed with SEU, LUG, AP1, and SVP. Upon repeated testing, however, WDR55 did not interact with any of these proteins (data not shown). This could be due to weak interactions, and thus not detectable in our Y2H system, or WDR55 does not directly interact on a protein level with these *AG* regulators.

Taken together, our data suggest a role of WDR55 in floral development. In particular it seems to control the pattern of *AG* expression independently from LUG-SEU repressor complex, indicating an additional pathway by which SVP repress *AG* expression. However, the function of *WDR55* in flowers does not seem to be restricted to the regulation of the boundaries of *AG* expression as exemplified by the *ag-1 wdr55-2* double mutant (Additional data file 1, Figure S8).

Discussion

The MADS-domain factor SVP has different functions during development. An 'early' function as a repressor of the floral transition and a 'later' function in floral meristem identity specification [6,8,18,20,24,25,48]. These two functions are also reflected by *SVP* expression, which is present in the leaves and SAM during the vegetative phase, is repressed in the meristem when plants switch to reproductive development and then reappears in the floral meristem during the early stages of flower development [8,24]. Whether SVP regulates different or similar sets of genes during these two phases of development is unknown. We employed ChIP-seq analysis to study the genome-wide binding behavior of SVP during these phases. SVP was found to bind to approximately 3,000 genes at both stages of development. Some genes were regulated by SVP at both stages of development, such as those in pathways regulating meristem development, whereas others were specific to one of the stages. One mechanism by which these differences in target gene specificity are likely to occur is through interactions between SVP and other MADS domain protein partners generating complexes with different specificities. Consistent with this idea, comparison of the targets of SVP and two of its partners, AP1 and FLC, showed similarities and differences.

Genome-wide ChIP-Seq experiments reveal several roles for SVP in modulating vegetative development

SVP bound to approximately 3,000 genes during vegetative development. GO terms analysis of these genes identified functional categories such as 'reproduction' and 'flower development' as being significantly over-represented in the list of putative SVP targets (Figure 3c). Similar results were previously found by Tao *et al.* [37]. These authors performed ChIP-chip experiments and identified a total of 328 genes bound by SVP during floral transition [37]. Comparison of the SVP target list of Tao *et al.* [37] and the list of targets of SVP at the vegetative stage presented here showed that only 15 genes are in common between the two datasets (Additional data file 7, Table S11). This discrepancy might occur for several reasons. First, Tao *et al.* made use of hybridization to Tiling arrays (ChIP-chip) to identify the genomic regions bound by SVP whereas in the present study these regions were identified by direct sequencing. As described previously, the set of peaks identified by the two technologies can be significantly different [74]. Second, in the ChIP-chip experiments of Tao and collaborators [37]*SVP* was expressed from the constitutive *CaMV35S* promoter whereas for the experiments shown here *SVP-GFP* was expressed from the native *SVP* promoter. MADS-domain transcription factors (including SVP) are expressed in specific tissues and interact with different partners to bind DNA in a tissue-specific manner [18], so the ectopic expression of SVP in all plant tissues and cell-types, as in *35S::SVP* plants, may affect the detection of the binding of this protein to genomic regions in a cell-specific context. Third, Tao *et al.* [37] identified SVP targets in 9-day-old seedlings grown under LDs. In the current study the vegetative tissue was harvested from *SVP::SVP-GFP svp-41* plants grown for 2 weeks under SDs (see Material and Methods). SVP interacting proteins might be expressed differently under these two conditions and therefore affect the capacity and/or selectivity of SVP to bind certain genomic regions.

Previously SVP was shown to delay flowering by directly repressing transcription of *FT* and *SOC1*, and reducing the mRNA level of the *FT* paralogue *TSF* [6,18,19]. Here, direct binding to *TSF* was not detected suggesting SVP might repress its transcription indirectly. FT and TSF are components of the photoperiodic flowering pathway, while *SOC1* is activated by FT in the SAM and acts as a point of convergence of other pathways [75-77]. Analysis of the flowering-time genes present in the high confidence list of SVP targets in vegetative tissue detected other genes acting in the photoperiodic flowering pathway or in the circadian clock that acts upstream of it. Notably, *GI* and *PRR7* are targets of SVP and both are involved in the photoperiodic induction of flowering and circadian clock regulation [78-80]. Both genes are positive regulators of *CO*, which in turn activates *FT* transcription under long photoperiods. Also the increase in SVP protein accumulation in the *lhy cca1* double mutant in continuous light, points to a link between SVP regulation and light/clock signaling [55].

The ChIP-seq data suggest that SVP likely also affects flowering by other mechanisms. The *FT* gene is a target for PRC2 and carries the chromatin mark H3K27me3 [81,82]. Therefore the regulation of PRC2 components by SVP may have an indirect effect on *FT* expression.

Mutations in components of PRC2, such as *CLF* that was also identified as a SVP target, cause ectopic expression of MADS-domain proteins that can then promote earlier flowering by mechanisms that remain unclear [83]. Furthermore, PRC2 and other chromatin-related targets of SVP reduce the expression of *FLC* [84], which encodes another MADS-domain protein that is a strong repressor of flowering and physically interacts with SVP [18,55,85]. This complex of FLC and SVP also binds directly to SVP, as discussed later, likely leading to repression of *SVP* transcription. Thus SVP appears to influence flowering time through several pathways that include chromatin regulation and feedback regulation on its own expression, as well as direct binding to genes encoding components of the circadian clock, photoperiodic flowering pathway and floral integrators.

SVP binds to genes involved in hormonal pathways

Our ChIP-seq data revealed numerous putative direct targets of SVP involved in hormonal pathways. SVP binds to genes involved in auxin, GA, cytokinin, and jasmonate homeostasis (Additional data file 3, Table S4). One of these direct targets is *STIP*, a gene involved in the maintenance of the pluripotency and proliferation of meristematic tissue in *Arabidopsis* [86]. Overexpression of *STIP* was shown to partially restore the SAM of the cytokinin insensitive *ahk2-2 ahk3-3 cre1-12* triple mutants, indicating that *STIP* acts downstream of CKs in the establishment of the SAM during early seedling development [49]. Several studies detected a role for cytokinins in the promotion of the floral transition [87]. For instance, the mutant *altered meristem program 1* (*amp1*) contains elevated levels of cytokinins and flowers earlier than wild-type plants [88]. Interestingly, the *amp1* mutant rescues the late-flowering phenotype of the *gi* mutant, demonstrating that CK is implicated in the LD pathway downstream of *GI* [50]. Our qRT-PCR experiments showed that *STIP* mRNA is induced in *svp-41* and in *ft-10 tsf-1 svp-41* (Figure 6a, b). This result indicates that SVP represses *STIP* independently or downstream of the two major photoperiod outputs *FT* and *TSF*. In addition, the induction of *STIP* in *svp-41* correlates with increased mRNA expression of several cytokinin response genes, belonging to the *type-A ARRs* and *CRFs* transcription factor families (Figure 6c), in agreement with the proposed role of *STIP* in the CK signaling pathway [49]. Moreover, a significant number of genes deregulated in *svp-41* were also found to be differentially expressed in response to BA (Figure 6d). These results suggest that in the *svp-41* mutant the up-regulation of *STIP* leads to the activation of the CK signaling pathway.

Additional targets of SVP encode hormonal receptors such as *COI1* that may also explain changes in gene expression of signaling components of jasmonate (*JAZs* genes). Furthermore the auxin responsive genes SAURs increase in expression in *svp-41* mutants, and these changes may be caused by altered auxin signaling, as SVP binds directly to genes related to auxin transport, such as *BIG* [51]. These effects suggest that the developmental role of SVP is likely to involve complex regulation of hormonal signaling pathways.

Common targets of the dimerizing MADS-box factors FLC and SVP

MADS-box factors form multimeric complexes that are proposed to be important in determining their DNA binding specificity and thereby their function [15,89]. SVP interacts with FLC and they are proposed to repress flowering as part of a complex that binds to the *SOC1* and *FT* genes [6,18,55,77]. To determine how extensive the overlap in target genes between FLC and SVP is, we compared the vegetative SVP ChIP-seq dataset with the one recently published for FLC [32]. The 112 genes in common between FLC and SVP high confidence targets included *CYTOKININ RESPONSE 1* (*CRE1*/*CHASE*), supporting a role for both proteins in regulating cytokinin signaling, as discussed above for SVP. However, the ChIP-seq and ChIP-qPCR experiments suggest that SVP and FLC bind to different regions of the gene, with SVP binding in an exon and FLC in the promoter. By contrast, SVP and FLC bound to the same region on the *SVP* promoter suggesting that the heterodimer composed of SVP and FLC could control SVP expression by means of a feedback loop. Taken together this comparison suggests that FLC and SVP do bind to many genes in similar positions, supporting the idea that they often bind to targets as a heterodimer, however some targets appear to be bound by only one of the proteins, indicating that they also have unique targets. Such a conclusion is consistent with the genetic data, which demonstrated that *svp flc* double mutants flower earlier than either single mutant [18,55].

SVP is linked to meristem function during two phases of development

Analysis of the subset of SVP targets that is common to vegetative and reproductive development showed an enrichment of genes involved in meristem function. During vegetative development the SAM continuously produces new cells that sustain plant growth by producing leaves and lateral branches, whereas after its formation the FM enlarges in an undifferentiated state until late stage 2, after which floral organ formation is initiated. *WUS* has a central role in development of both of these stages, participating in the maintenance of the vegetative, inflorescence, and floral meristems [59]. The ChIP-seq analysis showed that SVP binds to regulators of different stages of meristem development and some of these converge on the regulation of *WUS*. The *WUS* expression domain is restricted to a small group of L3 cells in the center of the meristem by the action of the *CLAVATA* (*CLV*) genes [57]. Our data show that SVP binds *CLV1* in both vegetative and reproductive tissues and *CLV2* in vegetative tissue. Besides the CLAVATA pathway, other genes that restrict WUS expression, for instance HD-ZIPIII and *SPLAYED* (*SYD*) [58,90] are also targets of SVP. In vegetative tissues SVP binds four of the five HD-ZIPIII genes described in Arabidopsis, *PHB*, *PHV*, *REV*, and *ATHB8*, and during flower development SVP binds *PHB*. Interestingly, we observed that the patterns of expression of *CLV1* and *PHB* become broader in the SAM of *svp-41* mutants compared to Col-0 (Figure 7). These data suggest that SVP influences meristem development by directly binding to genes that act at different levels in the regulatory hierarchy. *SVP* mRNA abundance in the SAM falls as it undergoes conversion from a vegetative to an inflorescence meristem and this correlates with the meristem becoming more domed and increasing in size [18,19]. Reduced activity of SVP in the inflorescence meristem might therefore alter the activity of meristem maintenance pathways to compensate for size differences between the vegetative and inflorescence meristem.

Similarly, floral meristem activity is under control of the MADS-box gene *AG*, which represses *WUS* expression after stage 6 of flower development [91]. SVP and AP1 both repress *AG* expression in the floral meristem, which in turn prevents the repressive activity of *AG* on *WUS*. Interestingly, our data show that SVP control *CLV1* activity since it binds directly to its locus, in the *svp-41 agl24 ap1-12* triple mutant *CLV1* is upregulated (Figure 8a) and the induction of SVP-GR result in the downregulation of *CLV1*; however the pattern of *CLV1* expression is retained (Figure 7 k and l) suggesting a direct role of SVP in the regulation of *CLV1* mRNA quantity, but not in the spatial boundary. Since CLV1 is also involved in repressing *WUS* activity, the deregulation of *CLV1* could be the cause of the downregulation of *WUS* expression that we detected by *in situ* (Figure 7m, n). Together these data show that SVP and AP1 secure *WUS* expression in the floral meristem via two pathways: the direct repression of *AG* and through direct repression of *CLV1*. This hypothesis is further strengthened by the observation that in the *svp-41 agl24 ap1-12* triple mutant a reduction in floral organ number was observed [25], which is probably due to a decrease in meristem size resulting from increased *CLV1* activity. Indeed the analysis of floral meristem size that we performed in this study revealed that in the triple mutant the FMs are smaller compared to the wild-type (Figure 7q and Table 1) indicating a direct correlation between SVP action and different WUS regulatory pathways.

Common targets of AP1 and SVP

SVP together with AGL24 and AP1 controls floral meristem identity and these proteins are important to prevent early expression of floral homeotic genes, such as *AP3*, *PI*, *SEP3*, and *AG* in the floral meristem [20]. This repression of floral organ identity genes involves recruitment of the LUG-SEU repressor complex by the AP1-SVP heterodimer [25]. As soon as the sepal primordia start to differentiate from the FM *SVP* expression disappears, probably due to interaction between AP1 and SEP3, as the latter starts to be expressed during late stage 2 of flower development [92]. Comparison of the gene lists obtained by ChIP-seq experiments for SVP and AP1 [31] identified a significant number of common target genes. Since SVP is strictly expressed in the floral meristem (stages 1 and 2 of flower development), many of these common targets are likely regulated during FM formation rather than specification of floral organ identity. Notably among these common targets transcription factors are enriched. These transcription factors include those involved in meristem maintenance and development. *PHB*, *KAN1*, and *ARF3* are all bound by both SVP and AP1 and are upregulated in *svp-41 agl24 ap1-12* inflorescences and the induction of SVP-GR result in the downregulation of *PHB*, *KAN1*, and *ARF3* suggesting that SVP modulate their activity. *PHB*, *KAN1*, and *ARF3* are involved in the regulation of meristem development and floral organ formation [58,93-95]. Interestingly the activity of ARFs proteins is controlled by convergent auxin flow that is controlled by PIN proteins and SVP and AP1 bound the genomic region of *PIN1*, which is expressed in the IM as well as in the FM. Indeed the expression level of *PIN1* is repressed by SVP. Taken together, these data suggest that there are interactions between the different regulatory networks that control FM formation and differentiation.

Analysis of the SEP3 ChIP-seq dataset revealed that *CLV1*, *PHB*, *KAN1*, and *ARF3* are also bound by SEP3, which also interacts with AP1 [15]. The expression profiles of SVP and SEP3 are mutually exclusive, suggesting a different modulation of the expression of the same target genes by SVP and SEP3 during floral meristem specification and floral meristem differentiation.

SVP targets are enriched in post-transcriptional and post-translational regulators

Multiple layers of regulation of gene expression play important roles in plant development. Post-transcriptional regulation can enhance and extend the effects of transcriptional regulation. The observation that SVP targets are enriched in genes encoding post-transcriptional and post-translational regulators indicates that SVP may affect gene expression not only by directly binding to target genes and modulating their transcription, but also by indirectly influencing post-transcriptional regulation.

Protein ubiquitination influences the stability and localization of proteins, resulting in the modulation of their biological functions. Defects in ubiquitination pathways can result in abnormal floral organ identity as suggested by the functional analyses of the *DCAF1* and *CYP71* genes, which are part of Cullin-RING ubiquitin ligase complexes [67,96].

SVP binds to a large number of DCAF encoding genes in FMs suggesting that SVP could be involved in the control of both proteasome and epigenetically mediated regulation of floral processes (Additional data file 1, Table S9). Several SVP targets are linked to chromatin-mediated regulation, such as two uncharacterized WD40 proteins containing Bromodomains, known to bind acetylated lysine residues in histones [97]. Thus SVP likely controls developmental processes by regulating gene expression directly through transcriptional regulation and indirectly by modulating transcription of genes encoding post-transcriptional and post-translational regulators.

It was recently reported that the WDR protein WDR55 is a putative DCAF and may function in a CUL4 - DDB1^{WDR55} E3 ligase complex [70]. Interestingly we discovered that *WDR55* is a target of SVP, which bound its genomic locus in inflorescence tissues. Moreover *WDR55* results downregulated in *svp-41 agl24* and *svp-41 agl24 ap1-12* compared to the wild-type inflorescences indicating that SVP acts as a direct activator of *WDR55* expression in the floral meristem.

The role of WDR55 in floral organ ontogenesis

The analyses of the mutant *wdr-55-2* showed variable phenotype in flower development such as reduced number of organs, asymmetric and reduced sepal and petal size, and occasionally chimeric organs such as petaloid stamens and carpelloid stamen or sepals. *In-situ* hybridization analysis revealed that *AG* was misexpressed in the *wdr55-2* flower. In wild-type, *AG* expression is always restricted to the two inner whorls (whorls 3 and 4). In homozygous *wdr55-2* mutant flowers *AG* expression is detectable earlier than in wild-type and in all floral whorls. This strongly suggests that WDR55 is involved in both spatial and temporal regulation of *AG*. The SVP-AP1 heterodimer is thought to recruit LUG-SEU and regulate *AG* expression in early stages of flower development [25]. We tested if WDR55 could bind any of these proteins but were not able to show any interaction.

Taken together the overall data indicate that SVP repress *AG* expression through two different pathways, the first is via the interaction with the co-repressor complex containing LUG-SEU and the dimer SVP-AP1 [25] and the second by SVP controlling the expression level of *WDR55*. The floral phenotype of the *wdr55-2* mutant is variable and did not result in the deregulation of *AG* in all the flowers, this suggests that SVP in the *wdr55-2* background is, although less efficient, still able to repress *AG* directly probably via the LUG-SEU pathway.

Conclusions

In summary, our data indicate that the SVP genome-wide binding profiles during two distinct developmental stages show a significant overlap and that this subset of genes includes a wider set of important regulators of plant development than was previously realized. However, there is also a large group of SVP target genes that are not bound at both stages, clearly reflecting distinct functions during vegetative and reproductive phases. The specificity of SVP binding to DNA is probably influenced by interaction with different MADS-domain partners, such as FLC and AP1. A related observation was made for the Drosophila MADS domain protein MEF2 that is expressed widely during development, but has specific targets at different stages dependent on the presence of interacting transcription factors [98]. The presented data provide new insights into the enormous diversity of pathways that are regulated by SVP and forms a basis for detailed analysis of the roles of SVP in regulating specific genes and pathways in combination with different interacting proteins.

Materials and methods

Plant material and growth conditions

For ChIP and microarray analysis of vegetative phase, *SVP::SVP-GFP*, *svp-41* single mutant (for plasmid construction see [20]) and wild-type seedlings were grown 14 days under short-day (SD) conditions (8 h light/16 h dark) at 22°C. For ChIP and microarray analysis of the reproductive phase, *SVP::SVP-GFP svp-41*, triple mutant *svp-41 agl24-2 ap1-12* and wild-type plants were grown under long-day (LD) conditions (LD; 16 h light/8 h dark) at 22°C. For the GR induction study the triple mutant *svp-41 agl24-2 ap1-10* was used [24]. All the plants were from the same Columbia ecotype. The *SVP::SVP-GFP svp-41* transgenic line and triple mutant *svp-41 agl24-2 ap1-12* have been previously described [20,25]. *ft-10 tsf-1 svp-41* and *ft-10 tsf-1* were described previously in Jang *et al.* [19]. The *wdr55-2* (WiscDsLox430F06) line is in the Col-0 ecotype and is a T-DNA insertion mutant obtained from the Nottingham Arabidopsis Stock Centre [99]. Seeds were surface sterilized using EtOH, bleach and Tween20 before germinated on MS media [100] supplemented with 2% sucrose (MS-2) and glufosinate-ammonium for BASTA selection of *wdr55-2* plants. All seeds were stratified on MS-2 plates at 4°C O.N. before being transferred to 18°C for about 12 days until germination. The seedlings were eventually transferred to soil and grown at 18°C under LD conditions (16 h).

ChIP assays

For ChIP experiments, the commercial antibody GFP: Living Colors_ full-length A.v. polyclonal antibody was used (Clontech [101]). Chromatin was prepared from inflorescences (2 weeks after bolting) and from 14-day-old seedlings of *svp*, grown under SD conditions. Wild-type plants (inflorescences and seedlings) were used as negative controls. ChIP assays were performed as previously described by [20] and in Additional data file 1, Methods S1 with a minor modification in the sonication step. DNA samples were sonicated six times 30 s each with amplitude 30 to 40, with intervals of 1 min (100-500 bp range fragments obtained).

We used as a positive control for the ChIP in the reproductive phase a region of the *AG* second intron (AG.V) that previously has been demonstrated to bind SVP-GFP [20]. For the vegetative phase we used regions in *FT* bound by SVP [18] (Additional data file 1, Figure S1). Enrichment fold to evaluate the quality of each ChIP sample was tested by qRT-PCR as described in Additional data file 1, Methods S2, all the primers used for ChIP-qPCR are in Additional data file 1, Table S12).

Sample preparation for ChIP-seq Illumina/Solexa sequencing

Two independent ChIP experiments (enrichment fold controlled by real-time PCR) were used for vegetative and reproductive ChIP-seq assays, respectively. We used one ChIP DNA sample for each library preparation and these were run on the Genome Analyzer. The DNA quantification of immunoprecipitated DNA was performed with the Quant-iT dsDNA HS Assay Kit (Invitrogen). Libraries for Solexa sequencing were prepared following the Illumina kit protocol, with some modifications. The first step 'Perform End Repair' was repeated twice, adding fresh enzymes and incubating 1 h longer than indicated by the protocol. Two units of undiluted Klenow enzyme was used. The incubation time of the step 'Ligate adapters to DNA fragments' was prolonged to 1 h instead of 15 min. Each library was validated quantifying the DNA with Quant-iT dsDNA HS Assay Kit (Invitrogen).

Read mapping and identification of enriched regions

Sequence reads were mapped to the unmasked Arabidopsis genome (TAIR8 build) using the Seqmap tool [102], allowing at most two mismatches at any position. Trimming unmapped reads at the 5' or 3' end led to marginal improvements in the number of reads mapped, and this step was therefore skipped. Reads belonging to duplicate experiments in each of the three conditions were pooled together. Only reads mapping to a unique position on the genome were considered for further analysis. This resulted in about 3 million uniquely mapped reads for the two inflorescences experiments, 5 million for seedlings experiments, and 6 million for control experiments. In each experiment, uniquely mapped reads were extended by 300 bps along the 5'->3' direction. This resulted in a base pair by base pair coverage map of the genome, that is, giving for each base pair the number of extended sequence reads that contained it. Only base pairs covered by reads mapping on both strands were considered valid for further analysis. Enrichment was then calculated in each valid base pair by comparing, for each IP experiment, the coverage in the experiment to the coverage in the control used as expected value, and computing an enrichment *P* value with a negative binomial distribution. In each comparison, the coverage of the two samples was normalized according to the number of reads obtained in each. Enriched regions were then defined as regions consisting of consecutive base pairs characterized by calculated *P* values <0.01 and not interrupted by a gap of 100 or more base pairs that were either non-valid or with a *P* value >0.01. The *P* value associated with each of these regions was defined as the minimum *P* value among the base pairs belonging to the region. Regions <150 bps were then discarded regardless of the *P* value. The number of remaining candidate-enriched regions was finally used to compute a Bonferroni corrected *P* value to be associated to the regions themselves. The overall strategy we followed in our analysis for the identification of enriched regions is highly similar to the one adopted in the SEP3 and AP1 ChIP-Seq experiments [13,31] and in the CSAR peak-finding tool [41], which has been shown to be better suited for ChIP-Seq experiments in Arabidopsis. *P* values for enrichment were computed by using a negative binomial distribution instead of the Poisson, as the former provides a better fit to count data from ChIP-Seq experiments [103]. Also, we employed a more conservative Bonferroni correction for multiple testing aimed at minimizing the number of false positive predictions.

Starting from regions with corrected *P* values <0.01, potential target genes were then identified by associating with each gene an overall *P* value given by the product of the *P* values associated with the single binding regions located in its gene locus, from 3 kbps upstream of the transcription start site to 1 kbp downstream of the transcribed region. Protocols of ChIP, DNA extraction, sequencing preparation, data processing, and all the associated files to this study can be found in the GEO (Gene Expression Omnibus) database (ID: GSE33120).

Tiling array experiments

The vegetative tissue samples were obtained from aerial parts of the *svp-41* single mutant and wild-type seedlings grown for 2 weeks under SD conditions (8 h light/16 h dark) and harvested at zeitgeber 8 (ZT8). For the reproductive tissue sampling we used wild-type and *svp-41 agl24-2 ap1-12* triple mutant inflorescences grown for 2 weeks under SD conditions and then moved to LD conditions (16 h light/8 h dark). The inflorescences were collected at 2 weeks after bolting at ZT8. RNA from three independent biological replicates was extracted using the RNA Plant Mini kit, QIAGEN (www1.qiagen.com/) and quantified by NanoDrop; 1 μg of total RNA was reverse transcribed into cDNA using an oligo(dT)-T7 primer, and was then converted into cRNA and linearly amplified by T7 *in-vitro* transcription reaction using the standard Ambion protocol (MessageAmp aRNA Kit, Ambion). cRNA was then reverse transcribed with random primers to dUTP-containing ds cDNA (WT ds cDNA Synthesis Kit, catalog no. 900813; Affymetrix). Fragmentation and labeling was performed with the GeneChip WT double-stranded DNA Terminal Labeling Kit (catalog no. 900812, Affymetrix). After fragmentation, 7.5 ug of ds-cDNA was hybridized for 16 h at 45°C on GeneChip Arabidopsis Tiling 1.0R Array. GeneChips were washed and stained with Fluidics Script FS450_0001 in the Affymetrix Fluidics Station 450. Then, the GeneChips were scanned using the GeneChip Scanner 3000 7G. Data were processed in R as described in [104]. Probe-level data were pre-processed using the RMA algorithm implemented in the Bioconductor package Affy. Linear models and empirical Bayes methods from the Limma package of Bioconductor were applied to derive a *P* value, false discovery rate (FDR; P adjusted), and mean of log2-based ratio across replicates. The data were deposited in the GEO (Gene Expression Omnibus) database (ID: GSE32397).

Gene Ontology analysis

The Bingo 2.44 plug-in [105] implemented in Cytoscape v2.81 [106] was used to determine and visualize the GO enrichment according to the GOslim categorization. A hypergeometric distribution statistical testing method was applied to determinate the enriched genes and the Benjamini and Hochberg FDR correction was performed in order to limit the number of false positives. The FDR was set up to 0.001 and 0.05 for the ChIP-seq and expression data, respectively. In addition to Bingo 2.44, further GO annotation analysis of the targets of SVP was performed by using TAIR bioinformatics resources [107].

cDNA preparation and qRT-PCR analysis

Expression analyses in the vegetative phase was performed using the *svp-41* single mutant, *35S::SVP* and wild-type seedlings grown for 2 weeks under SD conditions; for the reproductive phase we used wild-type and *svp-41 agl24-2 ap1-12* triple mutant inflorescences grown for 2 weeks under SD conditions and then moved to LD conditions. The inflorescences were collected at 2 weeks after bolting.

Total RNA from three biological replicates was extracted with the LiCl method, and its integrity was checked on agarose gels. The samples were treated with DNase (TURBO DNA-free; Ambion [108]) and reverse transcribed according to the ImProm-II_ Reverse Transcription System (Promega [109]) instructions. Sequence primers for RT-PCR amplification are listed in Additional data file 1, Table S13. Ten-fold dilutions of cDNA were tested in RT-PCR and qRT-PCR experiments using reference genes.

Enrichment folds were detected using a SYBR Green assay (Bio-Rad [110]). The real-time PCR assay was performed in triplicate using a Bio-Rad C1000 Thermal Cycler optical system or LightCycler480 (ROCHE) thermal cycler. For expression analyses normalized expression was calculated using the delta-delta Ct method (DDC(t)). For ChIP experiments, relative enrichment was calculated as described in Additional data file 1, Methods S2. For the expression analysis ubiquitin, PEX4, and PP2a-F were used as reference genes.

In-situ hybridization

In-situ hybridization has been performed as described in Additional data file 1, Method S3. The *WUS* antisense probe has been cloned according to Brambilla *et al.* [111]. The *ARF3* antisense probe has been cloned in the pGEM-T easy using the primers FW-CCCATCTGTATCAT-CATCACC and REV- CTCTCATTGCATAGATGTCC. The *KAN1* antisense probe has been cloned in the pGEM-T easy using the primers FW- AAGACCACTAA-CAAGCCTGC and REV- CATTTCTCGTGCCAATC TGGTC. The *CLV1* antisense probe has been cloned according to Clark *et al.* [60]. The *PHB* antisense probe has been cloned in the pGEM-T easy using the primers FW-GGTAGCGATGGTGCAGAGG and REV- CGAAC-GACCAATTCACGAAC. Sections were observed using a Zeiss Axiophot D1 microscope (Zeiss [112]) equipped with differential interface contrast (DIC) optics. Images were captured on an Axiocam MRc5 camera (Zeiss) using the AXIOVISION program (version 4.4).

Scanning electron microscopy

SEM has been performed as described in Additional data file 1, Method S4.

Inducible expression experiments

The *p35S::SVP-GR* construct was produced as follows: the coding region of *SVP* was amplified from inflorescence cDNA using primers Fw-CGTTGCCATGGCGAGAGAA AAGAT and Rev- ATTGTTCGGATCCCCACCACCA-TACGG containing NcoI and BamHI sites, respectively, cloned into pGEM-T easy (Promega), digested with NcoI and BamHI and ligated into pBluescript SK (Stratagene) containing a portion of the rat glucocorticoid hormone binding domain (a.a 508-795 [61]) to produce *pSK-SVP-GR*. The *AG-GR* fragment was amplified from the *pSK-SVP-GR* using the primers For and Rev and subcloned into the pTOPO vector (Life Technology). Finally *SVP-GR* was subcloned into the Gateway destination vector pB2GW7.0 [113] containing the 35S promoter. *p35S::SVP-GR* was transformed in svp-41 agl24-2 ap1-10 background (*ap1-10* heterozygous) and the T1 generation was selected for BASTA resistance.

After bolting, inflorescences of *35S::SVP-GR svp-41 agl24-1 ap1-10* plants were treated with a solution containing 10 μM dexamethasone (Sigma-Aldrich), 0.01% (v/v) ethanol, and 0.015% (v/v) Silwet L-77. Mock treatment consist of 0.01% (v/v) ethanol, and 0.015% (v/v) Silwet L-77.

For each time point, tissue from eight plants was collected. Tissue was removed as close to the surface of the inflorescence as possible to ensure an enrichment of FM cells.

Appendix

Accession numbers

Arabidopsis Genome Initiative locus identifiers for the genes mentioned in this article are as follows: *AGL24* [TAIR:AT4G24540], *STK* [TAIR:AT4G09960], *AP3* [TAIR:AT3G54340], *FLC* [TAIR:AT5G10140], *SVP* [TAIR:AT2G22540], *JAZ6* [TAIR:AT1G72450], *AGL16* [TAIR:AT3G57230], *SOC1* [TAIR:AT2G45660], *CLV1* [TAIR:AT1G75820], *PIN1* [TAIR:AT1G73590], *ARF3/ETT* [TAIR:AT2G33860], *KAN1* [TAIR:AT5G16560], *PHB* [TAIR:AT2G34710], *JAZ7* [TAIR:AT2G34600], *SADHU* [TAIR:AT3G42658], *JAZ8* [TAIR:AT1G30135], *GA2ox6* [TAIR:AT1G02400], *ARR6* [TAIR:AT5G62920], *ARR7* [TAIR:AT1G19050], *DDF1* [TAIR:AT1G12610], *GA2ox2* [TAIR:AT1G30040], *miR167* [TAIR:AT1G31173], *ACD6* [TAIR:AT4G14400], *AP1* [TAIR:AT1G69120], WDR55 [TAIR:AT2G34260], *VRN2* [TAIR: AT4G16845], *CLF* [TAIR: AT2G23380], *SWN* [TAIR: AT4G02020], *GI* [TAIR: AT1G22770], *FLK* [TAIR: AT3G04610], *FLD* [TAIR: AT3G10390], PRR7 [TAIR: AT5G02810], *PHYA* [TAIR: AT1G09570], *STIP* [TAIR: AT2G33880], *ARR11* [TAIR: AT1G67710], *ARR5* [TAIR: AT3G48100], *ARR15* [TAIR: AT1G74890], *CRF2* [TAIR: AT4G23750], *CRF5* [TAIR: AT2G46310], *PHV* [TAIR: AT1G30490], *REV* [TAIR: AT5G60690], *ATHB8* [TAIR: AT4G32880], *ATBARD1* [TAIR: AT1G04020], *KAN2* [TAIR: AT1G32240], *LMI1* [TAIR: AT5G03790], *DCAF1* [TAIR: AT4G31160], *JAZ5* [TAIR: AT1G17380], *JAZ10* [TAIR: AT5G13220], *JAZ1* [TAIR: AT1G19180]

Additional material

Additional data file 1: contains: Figure S1: Analysis of chromatin sample used for ChIP-seq experiments. **Figure S2:** qRT-PCR validation of differentially expressed genes between Col-0 and *svp-41* plants at the vegetative phase. **Figure S3:** GO enrichment analysis of differentially expressed genes between Col-0 and *svp-41* plants at the vegetative stage. **Figure S4:** Venn diagram containing the overlapping set of putative targets between SVP and FLC and SVP, AP1, and SEP3. **Figure S5:** Biologically active SVP-GR fusion. **Figure S6:** Flower organs of in *wdr55-2* -/- mutants show reduced size and asymmetric positioning. **Figure S7:** *In-situ* hybridization of wild-type and *wdr55-2* inflorescence using *AP3* and *PI* probes. **Figure S8:** Flower morphology of *wdr55-2 ag-1* mutant. **Table S1:** Summary of sequencing and mapping. **Table S3:** List of putative targets of SVP related to flowering time. **Table S6:** List of genes differentially expressed in *svp-41* compare to *Col-0* and related to auxin, cytokinin, or jasmonate homeostasis. **Table S9:** List of WDxR motif containing proteins found in SVP DNA binding screen. **Table S10:** Flower organ count from *wdr55-2* -/- mutants. **Table S12:** Primer pairs used for ChIP-qPCR assays. **Table S13:** Primer pairs used for the qRT-PCR expression analysis. **Methods S1:** ChIP protocol. **Methods S2:** qRT-PCR. **Methods S3:** *In-situ* hybridization. **Methods S4:** Scanning electron microscopy.

Additional data file 2: contains Table S2: High confidence targets of SVP in vegetative and reproductive tissues; list of the targets of SVP bound in both vegetative and reproductive tissues; lists of binding regions of SVP in vegetative and reproductive tissues.

Additional data file 3: contains Table S4: Lists of putative SVP targets with annotated functions in: meristem development in vegetative and reproductive tissues; response to hormonal stimuli such as auxin, cytokinin, ethylene, abscisic acid, jasmonate, and gibberellins in vegetative tissue.

Additional data file 4: contains Table S5: Tiling array expression data obtained using RNA extracted from: wild-type Col-0 and *svp-41* plants at the vegetative stage, inflorescences of wild-type Col-0 and *svp-41 agl24 ap1-12* and overlap between tiling array and ChIP-seq data.

Additional data file 5: contains Table S7: Lists of differentially expressed genes in *svp-41* mutant and the available expression-profiling data of seedlings treated with the CK benzyladenine (BA).

Additional data file 6: contains Table S8: putative targets for both SVP and AP1 and putative targets for both SVP and SEP3.

Additional data file 7: contains Table S8: Comparison of the SVP target list of Tao *et al*. [37] and the list of high confidence targets of SVP in vegetative tissue presented in this study.

Abbreviations

CArG-boxes: MADS-domain factors binding consensus; ChIP-qPCR: chromatin immunoprecipitation followed by quantitative real-time polymerase chain reaction; ChIP-seq: chromatin immunoprecipitation combined with high throughput DNA sequencing; ChIP: chromatin immunoprecipitation; CK: cytokinin; CSAR: ChIP-seq analysis in R; FM: floral meristem; GA: gibberellin; GEO: Gene Expression Omnibus; GO: Gene Ontology; GR: glucocorticoid receptor; IM: inflorescence meristem; qRT-PCR: quantitative real-time polymerase chain reaction; SAM: shoot apical meristem; Y2H: Yeast-2-Hybrid

Authors' contributions

VG did the ChIP experiments, biological analysis, and writing; FA did the transcriptome analysis, biological analysis, and writing; AS did the ChIP-seq and writing; RFG, SS, JLM, ST, and KNB performed the biological analysis; FZ and GMP were responsible for bioinformatics; GP did the bioinformatics and writing; LC conducted the research design; PEG, GC, and MMK did the research design and writing. All authors read and approved the final manuscript.

Acknowledgements
We thank V. Grandi, Marco Passaro, and Virginia Borrelli (Università degli Studi di Milano), K. Kaufmann (Wageningen University and Research Center, Wageningen) for technical support. Bruno Huettel (Max Planck Institute, Cologne) for the tiling array experiments. The SEM analysis was conducted at Centro Interdipartimentale di Microscopia Avanzata (CIMA) with the help of Dr. E. Caporali. We thank also Peter Huijser (Max Planck Institute, Cologne) for providing *35S::SVP* seeds. The post-doctoral and PhD fellowships for VG, SS, and FZ are funded by the Università degli Studi di Milano. FA is the recipient of a Marie Curie Intra-European Fellowship for Career Development (PIEF-GA-2009-251839). KNB and PEG were supported by grants from the Norwegian Research Council to PEG (grants 183190/S10 and 214052/F20). JLM is the recipient of a Humboldt Research Fellowship for Postdoctoral Researchers. GC received a grant from ERA-NET PG within the BLOOMNET project. This work was supported by the FLOWER POWER project (ID AGRO-11 and Ref. n° 16976) of the Lombardia region, Italy.

Author details
[1]Department of Bioscience, Università degli Studi di Milano, Via Celoria 26, 20133 Milan, Italy. [2]Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany. [3]Department of Biosciences, University of Oslo, N-0316 Oslo, Norway. [4]Consiglio Nazionale delle Ricerche Istituto di Biofisica, 20133 Milan, Italy.

Received: 24 March 2013 Revised: 24 April 2013
Accepted: 11 June 2013 Published: 11 June 2013

References

1. Irish VF: **Patterning the flower.** *Dev Biol* 1999, **209**:211-220.
2. Jack T: **Molecular and genetic mechanisms of floral control.** *Plant Cell* 2004, , **Suppl 1:** S1-17.
3. Ma H: **To be, or not to be, a flower-control of floral meristem identity.** *Trends Genet* 1998, **14**:26-32.
4. Melzer S, Lens F, Gennen J, Vanneste S, Rohde A, Beeckman T: **Flowering-time genes modulate meristem determinacy and growth form in Arabidopsis thaliana.** *Nat Genet* 2008, **40**:1489-1492.
5. Andrés F, Coupland G: **The genetic basis of flowering responses to seasonal cues.** *Nat Rev Genet* 2012, **13**:627-639.
6. Lee JH, Yoo SJ, Park SH, Hwang I, Lee JS, Ahn JH: **Role of SVP in the control of flowering time by ambient temperature in Arabidopsis.** *Genes Dev* 2007, **21**:397-402.
7. Mouradov A, Cremer F, Coupland G: **Control of flowering time: interacting pathways as a basis for diversity.** *Plant Cell* 2002, , **Suppl 1:** S111-S130.
8. Hartmann U, Hohmann S, Nettesheim K, Wisman E, Saedler H, Huijser P: **Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis.** *Plant J* 2000, **21**:351-360.
9. Parenicová L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, Angenent GC, Colombo L: **Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world.** *Plant Cell* 2003, **15**:1538-1551.
10. de Folter S, Angenent GC: **Trans meets cis in MADS science.** *Trends Plant Sci* 2006, **11**:224-231.
11. Colombo M, Masiero S, Vanzulli S, Lardelli P, Kater MM, Colombo L: **AGL23, a type I MADS-box gene that controls female gametophyte and embryo development in Arabidopsis.** *Plant J* 2008, **6**:1037-1048.
12. Masiero S, Colombo L, Grini PE, Schnittger A, Kater MM: **The emerging importance of type I MADS box transcription factors for plant reproduction.** *Plant Cell* 2011, **23**:865-872.
13. Kaufmann K, Muino JM, Jauregui R, Airoldi CA, Smaczniak C, Krajewski P, Angenent GC: **Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower.** *PLoS Biol* 2009, **7**:e1000090.
14. Theissen G, Saedler H: **Plant biology. Floral quartets.** *Nature* 2001, **409**:469-471.
15. de Folter S, Immink RG, Kieffer M, Parenicová L, Henz SR, Weigel D, Busscher M, Kooiker M, Colombo L, Kater MM, Davies B, Angenent GC: **Comprehensive interaction map of the Arabidopsis MADS Box transcription factors.** *Plant Cell* 2005, **17**:1424-1433.
16. Messenguy F, Dubois E: **Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development.** *Gene* 2003, **316**:1-21.
17. Simonini S, Roig-Villanova I, Gregis V, Colombo B, Colombo L, Kater MM: **Basic pentacysteine proteins mediate MADS domain complex binding to the DNA for tissue-specific expression of target genes in Arabidopsis.** *Plant Cell* 2012, **24**:4163-4172.
18. Li D, Liu C, Shen L, Wu Y, Chen H, Robertson M, Helliwell CA, Ito T, Meyerowitz E, Yu H: **A repressor complex governs the integration of flowering signals in Arabidopsis.** *Dev Cell* 2008, **15**:110-120.
19. Jang S, Torti S, Coupland G: **Genetic and spatial interactions between FT, TSF and SVP during the early stages of floral induction in Arabidopsis.** *Plant J* 2009, **60**:614-625.
20. Gregis V, Sessa A, Dorca-Fornell C, Kater MM: **The Arabidopsis floral meristem identity genes AP1, AGL24 and SVP directly repress class B and C floral homeotic genes.** *Plant J* 2009, **60**:626-637.
21. Yu H, Xu Y, Tan EL, Kumar PP: **AGAMOUS-LIKE 24, a dosage-dependent mediator of the flowering signals.** *Proc Natl Acad Sci USA* 2002, **99**:16336-16341.
22. Michaels SD, Ditta G, Gustafson-Brown C, Pelaz S, Yanofsky M, Amasino RM: **AGL24 acts as a promoter of flowering in Arabidopsis and is positively regulated by vernalization.** *Plant J* 2003, **33**:867-874.
23. Liu C, Chen H, Er HL, Soo HM, Kumar PP, Han JH, Liou YC, Yu H: **Direct interaction of AGL24 and SOC1 integrates flowering signals in Arabidopsis.** *Development* 2008, **135**:1481-1491.
24. Gregis V, Sessa A, Colombo L, Kater MM: **AGAMOUS-LIKE24 and SHORT VEGETATIVE PHASE determine floral meristem identity in Arabidopsis.** *Plant J* 2008, **56**:891-902.
25. Gregis V, Sessa A, Colombo L, Kater MM: **AGL24, SHORT VEGETATIVE PHASE, and APETALA1 redundantly control AGAMOUS during early stages of flower development in Arabidopsis.** *Plant Cell* 2006, **18**:1373-1382.
26. Kempin SA, Savidge B, Yanofsky MF: **Molecular basis of the cauliflower phenotype in Arabidopsis.** *Science* 1995, **267**:522-525.
27. Johnson D S, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316**:1497-1502.
28. Robertson AG, Bilenky M, Tam A, Zhao Y, Zeng T, Thiessen N, Cezard T, Fejes AP, Wederell ED, Cullum R, Euskirchen G, Krzywinski M, Birol I, Snyder M, Hoodless PA, Hirst M, Marra MA, Jones SJ: **Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding.** *Genome Res* 2008, **18**:1906-1917.
29. Nielsen R, Pedersen TA, Hagenbeek D, Moulos P, Siersbaek R, Megens E, Denissov S, Børgesen M, Francoijs KJ, Mandrup S, Stunnenberg HG: **Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis.** *Genes Dev* 2008, **22**:2953-2967.
30. Wederell ED, Bilenky M, Cullum R, Thiessen N, Dagpinar M, Delaney A, Varhol R, Zhao Y, Zeng T, Bernier B, Ingham M, Hirst M, Robertson G, Marra MA, Jones S, Hoodless PA: **Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing.** *Nucleic Acids Res* 2008, **36**:4549-4564.
31. Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, Kumar V, Serrano-Mislata A, Madueño F, Krajewski P, Meyerowitz EM, Angenent GC, Riechmann JL: **Orchestration of floral initiation by APETALA1.** *Science* 2010, **328**:85-89.
32. Deng W, Ying H, Helliwell CA, Taylor JM, Peacock WJ, Dennis ES: **FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of Arabidopsis.** *Proc Natl Acad Sci USA* 2011, **108**:6680-6685.
33. Immink RG, Posé D, Ferrario S, Ott F, Kaufmann K, Valentim FL, de Folter S, van der Wal F, van Dijk AD, Schmid M, Angenent GC: **Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators.** *Plant Physiol* 2012, **160**:433-449.
34. Yant L, Mathieu J, Dinh TT, Ott F, Lanz C, Wollmann H, Chen X, Schmid M: **Orchestration of the floral transition and floral development in Arabidopsis by the bifunctional transcription factor APETALA2.** *Plant Cell* 2010, **22**:2156-2170.

35. Zheng Y, Ren N, Wang H, Stromberg AJ, Perry SE: **Global identification of targets of the Arabidopsis MADS domain protein AGAMOUS-Like15.** *Plant Cell* 2009, **21**:2563-2577.
36. Winter CM, Austin RS, Blanvillain-Baufume S, Reback MA, Monniaux M, Wu MF, Sang Y, Yamaguchi A, Yamaguchi N, Parker JE, Parcy F, Jensen ST, Li H, Wagner D: **LEAFY target genes reveal floral regulatory logic, cis motifs, and a link to biotic stimulus response.** *Dev Cell* 2011, **20**:430-443.
37. Tao Z, Shen L, Liu C, Liu L, Yan Y, Yu H: **Genome-wide identification of SOC1 and SVP targets during the floral transition in Arabidopsis.** *Plant J* 2012, **70**:549-561.
38. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC: **Green fluorescent protein as a marker for gene expression.** *Science* 1994, **263**:802-805.
39. Liu C, Zhou J, Bracha-Drori K, Yalovsky S, Ito T, Yu H: **Specification of Arabidopsis floral meristem identity by repression of flowering time genes.** *Development* 2007, **134**:1901-1910.
40. Smyth DR, Bowman JL, Meyerowitz EM: **Early flower development in Arabidopsis.** *Plant Cell* 1990, **2**:755-767.
41. Muiño JM, Kaufmann K, van Ham RC, Angenent GC, Krajewski P: **ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions.** *Plant Methods* 2011, **7**:11.
42. Liljegren SJ, Gustafson-Brown C, Pinyopich A, Ditta GS, Yanofsky MF: **Interactions among APETALA1, LEAFY, and TERMINAL FLOWER1 specify meristem fate.** *Plant Cell* 1999, **11**:1007-1018.
43. Shore P, Sharrocks AD: **The MADS-box family of transcription factors.** *Eur J Biochem* 1995, **229**:1-13.
44. Meyerowitz EM: **DNA-binding properties of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS.** *Nucleic Acids Res* 1996, **24**:3134-3141.
45. Causier BE, Davies B, Sharrocks AD: **DNA binding and dimerisation determinants of Antirrhinum majus MADS-box transcription factors.** *Nucleic Acids Res* 1998, **26**:5277-5287.
46. Tang W, Perry SE: **Binding site selection for the plant MADS domain protein AGL15: an *in vitro* and *in vivo* study.** *J Biol Chem* 2003, **278**:28154-28159.
47. Pavesi G, Mereghetti P, Mauri G, Pesole G: **Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes.** *Nucleic Acids Res* 2004, **34(Web Server)**:W566-570.
48. Liu C, Xi W, Shen L, Tan C, Yu H: **Regulation of floral patterning by flowering time genes.** *Dev Cell* 2009, **16**:711-722.
49. Skylar A, Hong F, Chory J, Weigel D, Wu X: **STIMPY mediates cytokinin signaling during shoot meristem establishment in Arabidopsis seedlings.** *Development* 2010, **137**:541-549.
50. Bernier G, Perilleux C: **A physiological overview of the genetics of flowering time control.** *Plant Biotechnol* 2005, **3**:3-16.
51. Gil P, Dewey E, Friml J, Zhao Y, Snowden KC, Putterill J, Palme K, Estelle M, Chory J: **BIG: a calossin-like protein required for polar auxin transport in Arabidopsis.** *Genes Dev* 2001, **15**:1985-1997.
52. Yamaguchi N, Suzuki M, Fukaki H, Morita-Terao M, Tasaka M, Komeda Y: **CRM1/BIG-mediated auxin action regulates Arabidopsis inflorescence development.** *Plant Cell Physiol* 2007, **48**:1275-1290.
53. Yan J, Zhang C, Gu M, Bai Z, Zhang W, Qi T, Cheng Z, Peng W, Luo H, Nan F, Wang Z, Xie D: **The Arabidopsis CORONATINE INSENSITIVE1 protein is a jasmonate receptor.** *Plant Cell* 2009, **21**:2220-2236.
54. Sheard LB, Tan X, Mao H, Withers J, Ben-Nissan G, Hinds TR, Kobayashi Y, Hsu FF, Sharon M, Browse J, He SY, Rizo J, Howe GA, Zheng N: **Jasmonate perception by inositol-phosphate-potentiated COI1-JAZ co-receptor.** *Nature* 2010, **468**:400-405.
55. Fujiwara S, Oda A, Yoshida R, Niinuma K, Miyata K, Tomozoe Y, Tajima T, Nakagawa M, Hayashi K, Coupland G, Mizoguchi T: **Circadian clock proteins LHY and CCA1 regulate SVP protein accumulation to control flowering in Arabidopsis.** *Plant Cell* 2008, **20**:2960-2971.
56. Kurihara Y, Matsui A, Hanada K, Kawashima M, Ishida J, Morosawa T, Tanaka M, Kaminuma E, Mochizuki Y, Matsushima A, Toyoda T, Shinozaki K, Seki M: **Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis.** *Proc Natl Acad Sci USA* 2009, **106**:2453-2458.
57. Schoof H, Lenhard M, Haecker A, Mayer KF, Jurgens G, Laux T: **The stem cell population of Arabidopsis shoot meristems in maintained by a regulatory loop between the CLAVATA and WUSCHEL genes.** *Cell* 2000, **100**:635-644.
58. Prigge MJ, Otsuga D, Alonso JM, Ecker JR, Drews GN, Clark SE: **Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in Arabidopsis development.** *Plant Cell* 2005, **17**:61-76.
59. Sablowski R: **Flowering and determinacy in Arabidopsis.** *J Exp Bot* 2007, **58**:899-907.
60. Clark SE, Williams RW, Meyerowitz EM: **The CLAVATA1 gene encodes a putative receptor kinase that controls shoot and floral meristem size in Arabidopsis.** *Cell* 1997, **89**:575-585.
61. Gómez-Mena C, de Folter S, Costa MM, Angenent GC, Sablowski R: **Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis.** *Development* 2005, **132**:429-438.
62. Schwechheimer C, Calderon Villalobos LI: **Cullin-containing E3 ubiquitin ligases in plant development.** *Curr Opin Plant Biol* 2004, **7**:677-686.
63. Dumbliauskas E, Lechner E, Jaciubek M, Berr A, Pazhouhandeh M, Alioua M, Cognat V, Brukhin V, Koncz C, Grossniklaus U, Molinier J, Genschik P: **The Arabidopsis CUL4-DDB1 complex interacts with MSI1 and is required to maintain MEDEA parental imprinting.** *EMBO J* 2011, **30**:731-743.
64. He YJ, McCall CM, Hu J, Zeng Y, Xiong Y: **DDB1 functions as a linker to recruit receptor WD40 proteins to CUL4-ROC1 ubiquitin ligases.** *Genes Dev* 2006, **20**:2949-2954.
65. Lee J-H, Terzaghi W, Gusmaroli G, Charron J-BF, Yoon H-J, Chen H, He YJ, Xiong Y, Deng XW: **Characterization of Arabidopsis and rice DWD proteins and their roles as substrate receptors for CUL4-RING E3 ubiquitin ligases.** *Plant Cell* 2008, **20**:152-167.
66. Jin J, Arias EE, Chen J, Harper JW, Walter JC: **A family of diverse Cul4-Ddb1-interacting proteins includes Cdt2, which is required for S phase destruction of the replication factor Cdt1.** *Mol Cell* 2006, **23**:709-721.
67. Zhang Y, Feng S, Chen F, Chen H, Wang J, McCall C, Xiong Y, Deng XW: **Arabidopsis DDB1-CUL4 ASSOCIATED FACTOR1 forms a nuclear E3 ubiquitin ligase with DDB1 and CUL4 that is involved in multiple plant developmental processes.** *Plant Cell* 2008, **20**:1437-1455.
68. Angers S, Li T, Yi X, MacCoss MJ, Moon RT, Zheng N: **Molecular architecture and assembly of the DDB1-CUL4A ubiquitin ligase machinery.** *Nature* 2006, **443**:590-593.
69. Pazhouhandeh M, Molinier J, Berr A, Genschik P: **MSI4/FVE interacts with CUL4-DDB1 and a PRC2-like complex to control epigenetic regulation of flowering time in Arabidopsis.** *Pro Natl Acad Sci USA* 2011, **108**:3430-3435.
70. Bjerkan KN, Jung-Roméo S, Jürgens G, Genschik P, Grini PE: **Arabidopsis WD repeat domain55 interacts with DNA damaged binding protein1 and is required for apical patterning in the embryo.** *Plant Cell* 2012, **24**:1013-1033.
71. Weigel D, Meyerowitz EM: **The ABCs of floral homeotic genes.** *Cell* 1994, **78**:203-209.
72. Liu Z, Meyerowitz EM: **LEUNIG regulates AGAMOUS expression in Arabidopsis flowers.** *Development* 1995, **121**:975-991.
73. Franks RG, Wang C, Levin JZ, Liu Z: **SEUSS, a member of a novel family of plant regulatory proteins, represses floral homeotic gene expression with LEUNIG.** *Development* 2002, **129**:253-263.
74. Ho JW, Bishop E, Karchenko PV, Negre N, White KP, Park P: **ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis.** *BMC Genomics* 2011, **12**:134.
75. Samach A, Onouchi H, Gold SE, Ditta GS, Schwarz-Sommer Z, Yanofsky MF, Coupland G: **Distinct roles of CONSTANS target genes in reproductive development of Arabidopsis.** *Science* 2000, **288**:1613-1616.
76. Borner R, Kampmann G, Chandler J, Gleissner R, Wisman E, Apel K, Melzer S: **A MADS domain gene involved in the transition to flowering in Arabidopsis.** *Plant J* 2000, **24**:591-599.
77. Searle I, He Y, Turck F, Vincent C, Fornara F, Kröber S, Amasino RA, Coupland G: **The transcription factor FLC confers a flowering response to vernalization by repressing meristem competence and systemic signaling in Arabidopsis.** *Genes Dev* 2006, **20**:898-912.
78. Sawa M, Kay SA: **GIGANTEA directly activates flowering locus T in Arabidopsis thaliana.** *Proc Natl Acad Sci USA* 2011, **108**:11698-11703.
79. Nakamichi N, Kita M, Niinuma K, Ito S, Yamashino T, Mizoguchi T, Mizuno T: **Arabidopsis clock-associated pseudo-response regulators PRR9, PRR7 and PRR5 coordinately and positively regulate flowering time through the canonical CONSTANS-dependent photoperiodic pathway.** *Plant Cell Physiol* 2007, **48**:822-832.
80. Fowler S, Lee K, Onouchi H, Samach A, Richardson K, Coupland G, Putterill J: **GIGANTEA: a circadian clock-controlled gene that regulates photoperiodic flowering in Arabidopsis and encodes a protein with**

several possible membrane-spanning domains. *EMBO J* 1999, **18**:4679-4688.

81. Adrian J, Farrona S, Reimer JJ, Albani MC, Coupland G, Turck F: **cis-Regulatory elements and chromatin state coordinately control temporal and spatial expression of FLOWERING LOCUS T in Arabidopsis.** *Plant Cell* 2010, **22**:1425-1440.
82. Farrona S, Thorpe FL, Adrian J, Dong X, Sarid-Krebs L, Goodrich J, Turck F: **Tissue-specific expression of FLOWERING LOCUS T in Arabidopsis is maintained independently of polycomb group protein repression.** *Plant Cell* 2011, **9**:3204-3214.
83. Goodrich J, Puangsomlee P, Martin M, Long D, Meyerowitz EM, Coupland G: **A Polycomb-group gene regulates homeotic gene expression in Arabidopsis.** *Nature* 1997, **386**:44-51.
84. Amasino RM, Michaels SD: **The timing of flowering.** *Plant Physiol* 2010, **154**:516-520.
85. Jiang D, Wang Y, Wang Y, He Y: **Repression of FLOWERING LOCUS C and FLOWERING LOCUS T by the Arabidopsis Polycomb repressive complex 2 components.** *PLoS One* 2008, **3**:e3404.
86. Wu X, Dabi T, Weigel D: **Requirement of homeobox gene STIMPY/WOX9 for Arabidopsis meristem growth and maintenance.** *Curr Biol* 2005, **15**:436.
87. Bernier GJ: **My favourite flowering image: the role of cytokinin as a flowering signal.** *J Exp Bot* 2011.
88. Chaudhury AM, Letham S, Craig S, Dennis ES: **amp1: A mutant with high cytokinin levels and altered embryonic pattern, faster vegetative growth, constitutive photomorphogenesis and precocious flowering.** *Plant J* 1993, **4**:907-916.
89. Immink RG, Tonaco IA, de Folter S, Shchennikova A, van Dijk AD, Busscher-Lange J, Borst JW, Angenent GC: **SEPALLATA3: the 'glue' for MADS box transcription factor complex formation.** *Genome Biol* 2009, **10**:R24.
90. Wagner D, Meyerowitz EM: **SPLAYED, a novel SWI/SNF ATPase homolog, controls reproductive development in Arabidopsis.** *Curr Biol* 2002, **12**:85-94.
91. Lenhard M, Bohnert A, Jürgens G, Laux T: **Termination of stem cell maintenance in Arabidopsis floral meristems by interactions between WUSCHEL and AGAMOUS.** *Cell* 2001, **105**:805-814.
92. Mandel MA, Yanofsky MF: **The Arabidopsis AGL9 MADS-box gene is expressed in young flower primordia.** *Sex Plant Reprod* 1998, **11**:22-28.
93. Kerstetter RA, Bollman K, Taylor RA, Bomblies K, Poethig RS: **KANADI regulates organ polarity in Arabidopsis.** *Nature* 2001, **411**:706-709.
94. McConnell JR, Emery J, Eshed Y, Bao N, Bowman J, Barton MK: **Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots.** *Nature* 2001, **411**:709-713.
95. Sessions A, Nemhauser JL, McColl A, Roe JL, Feldmann KA, Zambryski PC: **ETTIN patterns the Arabidopsis floral meristem and reproductive organs.** *Development* 1997, **124**:4481-4491.
96. Li H, He Z, Lu G, Lee SC, Alonso J, Ecker JR, Luan S: **A WD40 domain cyclophilin interacts with histone H3 and functions in gene repression and organogenesis in Arabidopsis.** *Plant Cell* 2007, **19**:2403-2416.
97. Zeng L, Zhou M-M: **Bromodomain: an acetyl-lysine binding domain.** *FEBS Letters* 2002, **513**:124-128.
98. Cunha PM, Sandmann T, Gustafson EH, Ciglar L, Eichenlaub MP, Furlong EE: **Combinatorial binding leads to diverse regulatory responses: Lmd is a tissue-specific modulator of Mef2 activity.** *PloS Genet* 2010, **6**:e1001014.
99. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, Gadrinab C, Heller C, Jeske A, Koesema E, Meyers CC, Parker H, Prednis L, Ansari Y, Choy N, Deen H, Geralt M, Hazari N, Hom E, Karnes M, Mulholland C, Ndubaku R, Schmidt I, Guzman P, Aguilar-Henonin L, Schmid M, *et al*: **Genome-Wide Insertional Mutagenesis of Arabidopsis thaliana.** *Science* 2003, **301**:653-657.
100. Murashige T, Skoog F: **A rewised medium for rapid growth and bioassayswith tobacco tissue cultures.** *Physiologia Plantarum* 1962, **15**:473-497.
101. [http://www.clontech.com/].
102. Jiang H, Wong WH: **SeqMap: mapping massive amount of oligonucleotides to the genome.** *Bioinformatics* 2008, **24**:2395-2396.
103. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nat Biotechnol* 2008, **11**:1293-1300.
104. Naouar N, Vandepoele K, Lammens T, Casneuf T, Zeller G, van Hummelen P, Weigel D, Rätsch G, Inzé D, Kuiper M, De Veylder L, Vuylsteke M: **Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes.** *Plant J* 2009, **57**:184-194.
105. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** *Bioinformatics* 2005, **21**:3448-3449.
106. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD: **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc* 2007, **2**:2366-2382.
107. [http://www.arabidopsis.org/].
108. [http://www.ambion.com/].
109. [http://www.promega.com].
110. [http://www.bio-rad.com].
111. Brambilla V, Battaglia R, Colombo M, Masiero S, Bencivenga S, Kater MM, Colombo L: **Genetic and molecular interactions between BELL1 and MADS box factors support ovule development in Arabidopsis.** *Plant Cell* 2007, **19**:2544-2556.
112. [http://www.zeiss.com/].
113. Karimi M, Inzé D, Depicker A: **GATEWAY vectors for Agrobacterium-mediated plant transformation.** *Trends Plant Sci* 2002, **7**:193-195.

doi:10.1186/gb-2013-14-6-r56
Cite this article as: Gregis *et al*.: **Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in *Arabidopsis*.** *Genome Biology* 2013 **14**:R56.

Mitochondrion 20 (2015) 13–21

Contents lists available at ScienceDirect

Mitochondrion

journal homepage: www.elsevier.com/locate/mito

Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity

Anna Maria D'Erchia [a], Anna Atlante [b], Gemma Gadaleta [a], Giulio Pavesi [c], Matteo Chiara [c], Caterina De Virgilio [a], Caterina Manzari [b], Francesca Mastropasqua [a], Gian Marco Prazzoli [c], Ernesto Picardi [a], Carmela Gissi [c], David Horner [c], Aurelio Reyes [d], Elisabetta Sbisà [e], Apollonia Tullo [e], Graziano Pesole [a,b,*]

[a] Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università degli Studi di Bari Aldo Moro, Via Orabona 4, Bari 70126, Italy
[b] Istituto di Biomembrane e Bioenergetica, CNR, via Amendola 165/A, Bari 70126, Italy
[c] Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, Milano 20133, Italy
[d] Mitochondrial Biology Unit, Medical Research Council, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 0XY, United Kingdom
[e] Istituto di Tecnologie Biomediche- Sede di Bari, CNR, Via Amendola 122/D, Bari 70126, Italy

ARTICLE INFO

Article history:
Received 26 June 2014
received in revised form 23 September 2014
accepted 29 October 2014
Available online 1 November 2014

Keywords:
Mitochondrial DNA
Bioinformatics tools
Mitochondrial activity
Nucleo-mitochondrial cross-talk
Whole-exome sequencing
Mitochondrial gene expression

ABSTRACT

Eukaryotic cells contain a population of mitochondria, variable in number and shape, which in turn contain multiple copies of a tiny compact genome (mtDNA) whose expression and function is strictly coordinated with the nuclear one. mtDNA copy number varies between different cell or tissues types, both in response to overall metabolic and bioenergetics demands and as a consequence or cause of specific pathological conditions. Here we present a novel and reliable methodology to assess the effective mtDNA copy number per diploid genome by investigating off-target reads obtained by whole-exome sequencing (WES) experiments. We also investigate whether and how mtDNA copy number correlates with mitochondrial mass, respiratory activity and expression levels. Analyzing six different tissues from three age- and sex-matched human individuals, we found a highly significant linear correlation between mtDNA copy number estimated by qPCR and the frequency of mtDNA off target WES reads. Furthermore, mtDNA copy number showed highly significant correlation with mitochondrial gene expression levels as measured by RNA-Seq as well as with mitochondrial mass and respiratory activity. Our methodology makes thus feasible, at a large scale, the investigation of mtDNA copy number in diverse cell-types, tissues and pathological conditions or in response to specific treatments.

1. Introduction

Mitochondria play a range of critical roles in the life of eukaryotic cells (Pesole et al., 2012). They are the commonly referred to as the "power-stations" of the cell—for their provision of ATP through oxidative phosphorylation (OXPHOS), but they are also responsible for the biosynthesis of numerous macromolecules (lipids, proteins and nucleic acids) and contribute to the regulation of apoptosis, cell proliferation and motility. Individual cells, tissues and organs have distinct metabolic profiles and energy demands, which can change in response to environmental stimuli, alteration of physiological status or the onset of pathological conditions (Kunz, 2003; Leary et al., 1998; Leverve and Fontaine, 2001; Pfeiffer et al., 2001). Variation in mitochondrial respiratory capacity between tissues is also related to mitochondrial function, protein composition and morphology (Benard et al., 2006; Johnson et al., 2007a, 2007b; Mootha et al., 2003; Pagliarini et al., 2008).

Mitochondria are endowed of their own genetic system (mtDNA), a legacy of the endosymbiotic event that can be considered to represent the origin of eukaryotes and which occurred some 1.5–2 billion years ago (Lane and Martin, 2010). In mammals, the mitochondrial genome (mtDNA), a double stranded circular macromolecule, of around 16.5 kbp in length, is uniparentally inherited from the mother. Numerous mitochondria, each with multiple copies of the mtDNA, are present in each cell and form a highly dynamic network, the result of continuous fusion and fission processes (Rafelski, 2013). The function and activity of mitochondria is strictly regulated through the coordinated expression of mitochondrial and nuclear genomes. The mtDNA copy number per diploid nuclear genome correlates with ATP production and can range between 1000 and 5000 (Bogenhagen and Clayton, 1974; Shmookler Reis and Goldstein, 1983). mtDNA copy number is specific to tissue types and developmental stages, is under strict regulation (Clay Montier et al., 2009; Moraes, 2001), apparently a reflection of differing energy requirements. Altered mtDNA copy number is associated with oxidative stress and with several pathological conditions, including neuromuscular diseases, cardiomyopathy, type 2 diabetes and cancer (Clay Montier et al., 2009).

* Corresponding author at: via Amendola 165/A, 70125 Bari, Italy. Tel.: +39 0805443588; fax: +39 0805443317.
E-mail address: graziano.pesole@uniba.it (G. Pesole).

http://dx.doi.org/10.1016/j.mito.2014.10.005
1567-7249/© 2014 Elsevier B.V. and Mitochondria Research Society. All rights reserved.

A comprehensive vision of tissue- and pathology-related variability in mtDNA copy number per cell is lacking and few studies have attempted to correlate mtDNA copy number with mitochondrial mass, respiratory activity or mitochondrial gene expression levels.

We previously demonstrated that off-target reads from human whole-exome sequencing (WES) can be used to assemble mitochondrial genomes (Picardi and Pesole, 2012). Here, we investigate if the relative abundance of mtDNA reads obtained by WES experiments may be a reliable indicator of the effective mtDNA copy number per diploid genome, and whether and how mtDNA copy number correlates with global mitochondrial gene expression levels measured by RNA-Seq.

In the current study, we have estimated mtDNA copy number by qPCR (Venegas et al., 2011) as well as by the proposed WES-derived protocol for liver, kidney, brain, lung, muscle and heart samples from three age- and sex-matched human individuals. All samples were also subjected to whole transcriptome sequencing as well as to citrate synthase and cytochrome oxidase assays.

A highly significant linear correlation between qPCR data and the frequency of mtDNA off target reads imply that, in addition to allowing the reconstruction of complete or nearly complete mitochondrial genomes (Picardi and Pesole, 2012), WES data permit accurate quantification of mtDNA copy number. Furthermore, mtDNA copy number showed highly significant correlation with mitochondrial gene expression levels as measured using RNA-Seq as well as with key functional data. The relative abundance of distinct mtDNA-derived transcripts was tissue-specific and highly replicable.

2. Materials and methods

2.1. Samples and nucleic acids extraction

Six different post-mortem human snap-frozen tissues (brain, liver, lung, striated muscle, kidney and heart) from three unrelated healthy Caucasian individuals (males, aged 47–54 years) were obtained from Cureline (South San Francisco, CA, USA).

The three sample IDs are S7/11, S12/12 and S13/12, and all sample details are reported in Supplemental Table S1.

DNA was purified using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions, quantified and qualitatively checked on NanoDrop 2000c (Thermo Fisher Scientific, USA).

Total RNA was purified using the RNeasy Plus Mini Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions. RNA quality was assessed on Agilent Bioanalyzer 2100, obtaining RIN (RNA Integrity Number) values ranging from 5 to 7, that were considered acceptable for RNA derived from post-mortem tissues.

2.2. Quantification of mtDNA content

Relative mtDNA copy number was measured by qPCR with SYBR detection using primers specific for the mitochondrial tRNA Leu$^{(UUR)}$ gene and the single copy nuclear ß-2-microglobulin (ß2M) gene (Venegas et al., 2011). All reactions were performed using 5 ng of total DNA as template on a ABI Prism 7000 sequence detector system (Applied Biosystems, Foster City, CA, USA), according to this two-step thermal cycling protocol: 50 °C for 2 min (UDG pre-treatment); 95 °C for 10 min (initial denaturation); 40 cycles at 95 °C for 15 s and 62 °C for 1 min, followed by a melting curve analysis (95 °C for 15 s, 60 °C for 30s and 95 °C for 15 s) to verify the specificity and identity of the PCR product. Primers sequence, amplicon size and annealing temperature are reported in the Supplemental Table S2.

The intensity of SYBR fluorescent signals were then analyzed by the SDS software (version 1.2.3), and the C_T value for each qPCR was used to calculate the mtDNA content by difference in C_T values between the tRNA Leu$^{(UUR)}$ and ß2M genes (ΔC_T). mtDNA content was obtained using the formula $2 \times 2^{-(\Delta CT)}$. Results represent the average of three independent experiments performed on the same DNA preparation and are shown with standard deviations.

2.3. Exome sequencing and mtDNA assembly

Exome capture was performed using the TruSeq Exome Enrichment Kit (Illumina, San Diego, CA), according to the manufacturer's instructions. Briefly, for each tissue, a DNA library, including inserts ranging in size from 200 to 400 bp approximately, was prepared using the TruSeq DNA Sample Prep kit (Illumina). Then, each library was hybridated with biotinylated probes targeting the exonic regions (about 200,000 exons, covering about 62 Mb of the human genome). After two steps of enrichment with the probes, the captured exonic regions were sequenced on the Illumina HiSeq 2000 sequencer, at IGA Technology Services in Udine (Italy), generating for each tissue approximately 40 million of 100 bp paired-end reads.

Exome reads were mapped onto the Revised Cambridge Reference Sequence (rCRS with GenBank accession number NC_012920) of human mtDNA (Andrews et al., 1999) using GSNAP program version 2013-07-14 (Wu and Nacu, 2010) since it enables the handling of circular genomes.

Aligned reads were mapped again onto the complete human genome (assembly hg19 including the rCRS sequence) using GSNAP in order to exclude read pairs also mapping on nuclear mitochondrial DNA sequences (i.e., Numts) (for further details see Picardi and Pesole, 2012). The complete mapping procedure was automated by using a custom python script (mapExome.py), available upon request.

Mitochondrial reads in SAM format were converted into BAM and then pileup format using *samtools* (version 0.1.18) (Li et al., 2009). The final pileup file was parsed position-by-position in order to calculate the distribution of nucleotides aligned at each position, removing bases with a quality score less than 25. For each position of the rCRS sequence, supported by at least 5 independent reads, the consensus base was calculated using a minimum confidence level of 0.75.

Contiguous consensus positions were grouped in contigs and assembled into a final mitochondrial genome using a custom python script previously developed in our group (Picardi and Pesole, 2012).

The relative amounts of off-target mtDNA reads were calculated as the number of reads mapping on mtDNA per million mapped reads.

2.4. Strand-oriented RNA-Sequencing and analysis

For each tissue, a strand-oriented RNA library was prepared to preserve information about which DNA strand was the original template during the synthesis of transcripts, thus offering strand orientation for detection of antisense transcription and providing information about regulatory relationships.

The cytoplasmatic rRNA removal was performed for each total RNA sample using the Ribo-Zero rRNA removal Kit (Epicentre, Madison, WI, USA). The rRNA-depleted RNA was used to prepare the stranded-oriented RNA-seq library using the TruSeq Stranded Total RNA Sample Prep Kit (Illumina, San Diego, CA, USA), according to the manufacturer's instructions. Briefly, each RNA was chemically fragmented prior to the random priming reverse transcription reaction for first strand cDNA generation. The fragmentation step resulted in an RNA-seq library including inserts ranging in size from approximately 100–400 bp. During the second-strand synthesis, dUTP was incorporated in place of dTTP, thus preventing amplification of this strand during the subsequent PCR step and retaining strand information. cDNA libraries were sequenced on the Illumina HiSeq 2000 platform at IGA Technology Services in Udine (Italy), generating for each tissue sample approximately from 27 to 35 million 100 bp paired-end reads.

In order to investigate the expression level of mitochondrial-encoded transcripts, we used the reference annotation in Supplemental Table S3, which also includes long non-coding RNAs (Rackham et al., 2011).

No reads that could be attributed uniquely to mature tRNAs were mapped, as expected, as the procedure used for the RNA extraction does not provide an enrichment for transcripts <200 nucleotides long (such as miRNAs and tRNAs).

Sequences were first mapped against the human genome (hg19) using Tophat with default parameters, with a percentage of mapped reads in the different samples ranging from 88% to 94%. Sequence read pairs that mapped on mtDNA as a correct forward-reverse pair, with a maximum insert size of 1000 bp were considered to be initial mitochondrial candidate reads.

As expected, a non-negligible fraction of the latter (about 8% of the sequence pairs) resulted to be mapped on Numt regions of the nuclear genome as well, either as a singleton (only one of the two reads mapping), as a chimeric pair mapping on different nuclear chromosomes, or as a pair on the same chromosome. Since having a correct estimation of the number of sequence reads that could be reliably assigned to mitochondrial genes was essential for this study, we further investigated this issue. A comparison of matching against the mitochondrial and the nuclear genomes revealed that for more of the 90% of the ambiguously mapped sequence pairs a match was found for only one of the two paired reads on the nuclear genome. Only about 1% of the reads mapping as paired on the mitochondrial genome was mapped on Numts as correct forward–reverse pair, however with a number of mismatches equal to or greater than the one resulting from the mapping on the mitochondrion.

All in all, these results show how mapping ambiguity between the Numt sequences and the mitochondrion can be resolved by employing paired-end sequences. All reads mapping as a correctly oriented pair on the mitochondrial sequence can be considered for further analyses independently of mapping on the nuclear genome.

Starting from reads mapped on the mitochondrial genome, expression of genes was estimated by using the annotation summarized in Supplemental Table S3. Sequence reads were assigned to a gene when both sequence reads were completely contained within the gene boundaries and were assigned by directional sequencing to the same strand. Reads per kilobase per million (RPKM) values (Mortazavi et al., 2008) were then computed starting from these counts.

RNA–DNA variations were detected using REDItools (Picardi and Pesole, 2013).

2.5. Tissue homogenate preparation

The PBI-Shredder, an auxiliary high-resolution respirometry (HRR) Tool, was used to prepare homogenate—in 0.2 M phosphate buffer (pH 8.0)—of frozen tissue specimens (Draxl et al., 2013), with high reproducibility of mitochondrial function as evaluated with HRR by means of Oxygraph-2 k OROBOROS®. Homogenate protein content was determined according to (Waddell, 1956) with bovine serum albumin used as a standard.

2.6. Enzymatic activity measurements

Citrate synthase (CS) and cytochrome *c* oxidase (COX) activities were measured by spectrophotometric standard methods. Each assay was performed at least in triplicate by using homogenate tissues subjected to three freeze–thaw cycles to disrupt membranes and expose mitochondrial enzymes.

The reduction of 5,5-dithiobis(2-nitrobenzoic acid) (DTNB) by CS at 412 nm (extinction coefficient is 13.6 mM^{-1} cm^{-1}) was followed in a coupled reaction with coenzyme A and oxaloacetate (Robinson and Srere, 1985). A reaction mixture of 0.1 mM acetyl-coenzyme A, 0.2% Triton-x-100, 0.1 mM DTNB and 20–40 μg of homogenate protein was incubated at 25 °C for 5 min. The reaction was initiated by the addition of 0.5 mM oxaloacetate. Results are expressed as nanomoles CoA formed per minute per mg homogenate protein.

COX activity was measured by following the decrease of absorbance at 548–540 nm (extinction coefficient is 19.1 mM^{-1} cm^{-1}), due to the oxidation of 40 μM ferrocytochrome *c* (reduced with substoichiometric concentrations of potassium ascorbate), for 90 s at 25 °C (see (Bobba et al., 2013)). Very stringent controls, including (i) test with cyanide, (ii) adequate homogenate proteins used for the reaction and (iii) gel-filtration of the reduced cyt *c* (in a superfine Sephadex G-25 column) to remove both excess of reductant and dimeric/multimeric form of cyt *c*, which per se can inhibit the reaction itself were performed to assure correct estimates of the Cox assay. Time-dependent absorbance changes were recorded with a Jasco double-beam/double-wavelength spectrophotometer UV-550. The rate of ferrocytochrome *c* oxidation, obtained as tangents to the initial part of the progress curves, is expressed as nanomoles cyt c_{ox} formed per minute per mg homogenate protein.

Data were expressed as means ± standard deviation (S.D.) ($n > 3$) and analyzed with SPSS software by 1-way analysis of variance (ANOVA) for repeated measures followed by the *post hoc* Bonferroni test for multiple comparisons. Statistical differences were determined at $P < 0.05$.

3. Results

3.1. mtDNA copy number per diploid nuclear genome determined by qPCR

We measured the relative mtDNA copy number in six tissue types (brain, lung, kidney, liver, heart and skeletal muscle) from three unrelated individuals (S7/11, S12/12 and S13/12) by qPCR. The highest mtDNA copy number was observed in heart tissue, followed by skeletal muscle, brain, liver, kidney and lung (Fig. 1 and Supplemental Table S4) consistent with the hypothesis that tissues with higher ATP requirement should show higher mitochondrial copy number. Differences in mtDNA content between the three subjects are particularly evident in brain and muscle. The low mtDNA copy number observed in the brain individual S13/12 could be related to the cause of death (asphyxia), while the differences observed in muscle might reflect heterogeneous relative content of nuclei and mitochondria between muscle fibers or to different levels of habitual physical activity among the three sampled subjects.

3.2. Correlation between mtDNA copy number and unspecific enrichment of mtDNA reads

To evaluate whether the relative abundance of mtDNA reads among off-target exome enrichment reads was a reliable and effective estimator of mtDNA copy number, we generated and analyzed whole exome sequencing (WES) data for each of the 18 samples (Supplemental Table S5). The relative number of mtDNA reads among off-target WES reads was calculated using previously published methods (Picardi and Pesole, 2012). Bivariate linear correlation between mtDNA copy number and the relative amount of off-target mtDNA reads (Fig. 2) was highly significant ($r^2 = 0.92$, $P < 0.0001$) confirming the effectiveness and reliability of the proposed approach for quantifying mtDNA copy number.

3.3. Nucleotide variants in assembled mtDNA sequences

For each tissue and individual, we assembled a complete mtDNA from whole exome reads (see Methods). Analysis of mtDNAs did not show significant somatic variations between different tissues of the same individual. Using stringent criteria (minimal coverage of 50 reads and a base variation frequency higher than 10%), we found heteroplasmy higher than 90% at position 310 of the D-Loop (T-to-C) in two of the three individuals. These observations were supported by the corresponding RNA-Seq data (Supplemental Table S6). The stringent filters employed were imposed to minimize false-discovery rate

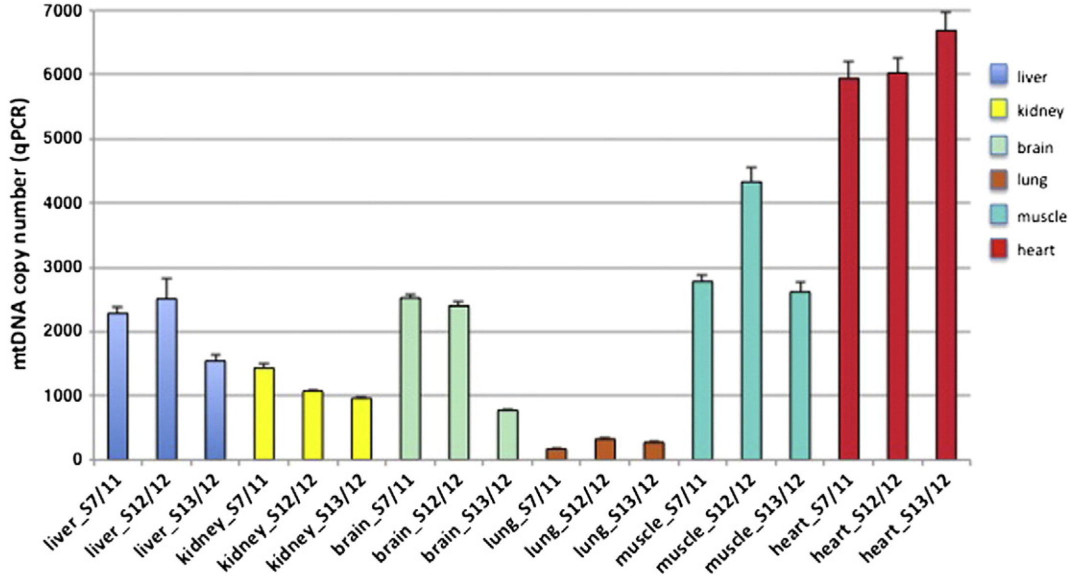


Fig. 1. Relative mtDNA copy number in six tissues of three individuals (S7/11, S12/12 and S13/12) calculated by qPCR. Values represent the average of three independent experiments.

but might lead to the exclusion of poorly supported heteroplasmic sites. Relaxing the variation frequency, we also detected low frequency (5%) heteroplasmy at position 310 in the third individual as well as an additional low-frequency heteroplasmic position at D-Loop position 72 (T-to-C) in one individual (Supplemental Table S6). Both of these D-Loop variant positions are annotated as heteroplasmic sites in MITOMAP (Ruiz-Pesini et al., 2007).

Sample matched comparisons between DNA (from whole exome) and RNA (from RNA-Seq) mitochondrial reads were used to identify potential post-transcriptional modification events. Only changes occurring at DNA homoplasmic sites with frequencies higher than 10% at the RNA level were considered. We observed tissue-specific nucleotide variations in four positions, three in heart and one in brain, consistently shared by all three individuals (Supplemental Table S7). Positions 1955 and 2617 (heart) were with the 16S rRNA, whereas site 905 (heart) fell in the 12S rRNA. Notably, at position 2617, we identified both A-to-U and A-to-G changes as previously reported (Bar-Yaacov et al., 2013). Position 8303 falls within the lysine tRNA and is found here for the first time as a target of a post-transcriptional modification event supported by RNA-Seq reads likely derived from polycistronic pre-processed transcripts which might include tRNAs (Nardelli et al., 1994). In this site both A-to-U and A-to-G changes were observed, but only in brain tissue. An additional RNA–DNA difference at position 295, described by Bar-Yaacov et al. (2013), was also observed in our samples, although not in all individuals and tissues because of the stringent filters employed (data not shown).

3.4. Determination of mitochondrial mass and tissue energy requirement

Citrate synthase (CS), a component of the tricarboxylic acid cycle, is a stably expressed mitochondrial matrix enzyme. Its specific activity is frequently used as an indicator of total mitochondrial mass (Figueiredo et al., 2008; Sarnat and Marín-García, 2005; Kirby et al., 2007). Cytochrome oxidase (COX) activity is often employed as a marker of OXPHOS activity since this enzyme (complex IV) constitutes the last step in the respiratory chain (RC), likely limiting its electron flux (Capaldi, 1990; Kunz et al., 2000; Larsen et al., 2012; Mazat et al., 2001; Villani and Attardi, 1997, 2000; Villani et al., 1998). To evaluate

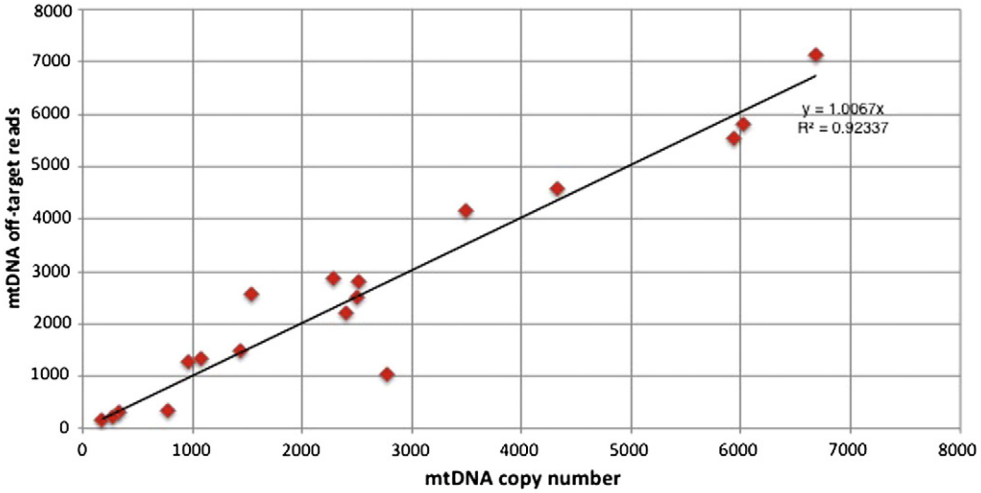


Fig. 2. Linear correlation between the mtDNA copy number in the six tissues of three individuals (S7/11, S12/12 and S13/12) and off-target mtDNA reads of whole exome sequencing data, calculated performing a bivariate linear fit analysis ($P < 0.0001$). The amount of off-target mtDNA reads has been calculated as the number of mtDNA reads per million WES reads (see Supplemental Table S5).

variation in mitochondria content and respiratory capacity, CS and COX activities were determined spectrophotometrically in total homogenates from all 18 samples.

Significant variation in CS and COX activities (normalized by the total homogenate protein content) were observed between tissues as well as between individuals (ANOVA). Tissue-dependent differences again reflected known differences in energetic demands between tissues with heart, the organ that consumes most energy per mass unit (Goffart et al., 2004; Van den Bogert et al., 1993), showing the highest CS and COX specific activities with values 4- to 12-fold and 1.2- to 2fold respectively higher than other tissues (Table 1).

Knowing the amount of nuclear DNA (nDNA) and mtDNA per mg of protein lysate of the same tissue specimens, it was possible to normalize the CS and COX enzymatic activities per nDNA and mtDNA. Table 1 shows that CS activity/nDNA ratio varies greatly between tissues, again reflecting the tissue-specific differences in mitochondrial activities. As expected, the heart shows the highest ratio in all three individuals, in accordance with mtDNA relative abundance values and evidence that it has the highest mitochondrial content per cell (Fernandez-Vizarra et al., 2011). Moreover, the CS activity/mtDNA also differs markedly between samples, with higher values for heart, skeletal muscle and brain (Table 1).

The highest COX activity/nDNA values were also associated with high energy-requirement tissues. However, COX activity/mtDNA values were, while variable, somewhat more consistent.

3.5. Correlation between mtDNA copy number and mitochondrial mass and respiratory capacity

We then evaluated if a correlation exists between the mtDNA copy number and the mitochondrial mass and the respiratory capacity in the different tissues. As shown in Fig. 3A and B, a positive and highly significant correlation was also found between either CS or COX activity per cell and the amount of mtDNA per cell. This indicates that the tissues which have the highest COX and CS activities have also the highest cellular mtDNA amount.

3.6. Mitochondrial gene expression in different tissues evaluated by RNA-seq

We carried out a strand-oriented RNA-Seq (2 × 100 bp paired-end, with random hexamer priming and no mitochondrial rRNA depletion) in the six tissue samples from the three individuals. Special attention was paid to minimize the impact of reads that could generate read mapping artifact to Numt (nuclear mitochondrial DNA) sequences (Calabrese et al., 2012) and introduce biases into expression level estimates (see Methods). Analysis of cDNA reads that mapped to Numts revealed that while rRNA like reads were most predominant, around 90% of ambiguously mapped read pairs from protein coding ORFs derived from CO1 or CO3 genes.

Globally, from 4% to 27% of read-pairs (according to tissue) were of mitochondrial origin (see % mtDNA PE reads in Supplemental Table S8 and Methods).

We observed a remarkably imbalanced level of expression between the mtDNA plus and minus strand, with about one thousand fold more reads mapping to the former. Furthermore, only about 0.36% of the reads mapping on the mtDNA plus strand resulted to be originated from the precursor RNA in contrast to over 85% of reads mapping on the minus strand (Supplemental Table S8 and Methods). Mapping of reads derived from the heavy strand covered the whole mitochondrion DNA sequence in all tissues (data not shown). The largest fraction of PE reads, around 95%, originated from 12S and 16S rRNA genes in all tissues considered. Other reads originated, as expected, mostly on annotated protein coding genes, with however sharp enrichment peaks also in the antisense strand of 16S rRNA, CO1 and ND5 genes (data not shown). Antisense transcripts of the ND5, ND6 and CytB genes have been demonstrated to correspond to three lncRNAs (Rackham et al., 2011), and we hypothesize that other antisense ncRNAs could originate in correspondence with the other peaks observed.

Expression levels of the 11 mtDNA protein coding mature transcripts, ribosomal RNAs, and lncRNAs are reported in Supplemental Table S9, expressed as RPKM values as well as relative expression within the same tissue. The comparison of absolute RPKM values of the same gene across different samples shows high variability, reflecting the different concentrations of mitochondrial RNAs in each sample. The relative expression of genes within the same sample, instead, remains remarkably constant across the different tissues (Fig. 4). However, we observe a considerable variability of this measure, corresponding to a highly variable level of steady-state expression of the different mature transcripts, suggesting a remarkable and variable effect of post-transcriptional cleavage, processing and stability mechanisms in the regulation of gene expression, that are conserved across the different tissues investigated.

3.7. Correlation of between mtDNA copy number and mitochondrial gene expression

Finally, we sought to investigate potential correlation between mtDNA copy number and the expression level of mtDNA-encoded genes. The overall concentration of RNAs of mitochondrial origin in

Table 1
Citrate synthase (CS) and cytochrome oxidase (COX)-specific activity measurements and normalization by mtDNA and nDNA content.

Tissue	CS-specific activity	COX-specific activity	CS/mtDNA	CS/nDNA	COX/mtDNA	COX/nDNA
liver_S7/11	530 ± 31	1357 ± 32	5196	27	13303	68
liver_S12/12	245 ± 14	1813 ± 47	4803	20	35549	144
liver_S13/12	297 ± 15	1411 ± 55	1650	7	7839	31
kidney_S7/11	367 ± 18	1260 ± 53	3336	13	11454	144
kidney_S12/12	246 ± 11	1442 ± 44	6648	22	38972	126
kidney_S13/12	426 ± 21	2832 ± 62	4260	11	28320	75
brain_S7/11	164 ± 10	949 ± 37	10933	71	63266	413
brain_S12/12	166 ± 11	1400 ± 43	6384	37	53846	318
brain_S13/12	166 ± 9	1153 ± 30	12769	22	88693	154
lung_S7/11	76 ± 5	1436 ± 41	4750	2	89750	36
lung_S12/12	62 ± 5	1522 ± 42	5166	4	126833	98
lung_S13/12	132 ± 7	1445 ± 47	2870	2	31413	21
muscle_S7/11	196 ± 10	995 ± 28	8521	75	43260	383
muscle_S12/12	475 ± 22	1363 ± 31	10106	131	29000	378
muscle_S13/12	335 ± 12	1335 ± 33	14565	86	58043	342
heart_S7/11	564 ± 28	1836 ± 41	16114	209	52457	680
heart_S12/12	768 ± 37	2639 ± 55	10378	183	35662	628
heart_S13/12	524 ± 25	1890 ± 47	12780	227	46097	821

Data are the mean + standard deviation, with $n = 4$, obtained from six human tissues, liver, kidney, brain, lung, muscle and heart, from three age- and sex-matching individuals.

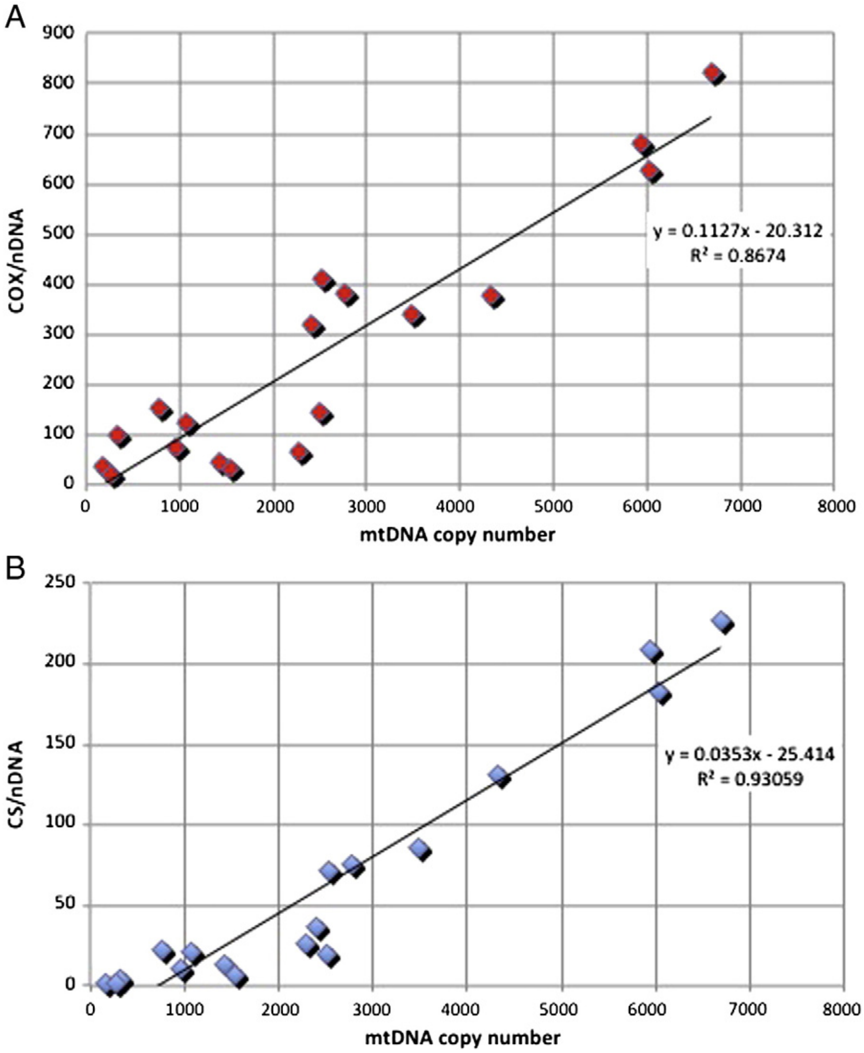


Fig. 3. Linear correlation between mtDNA copy number and mitochondrial mass and respiratory capacity. (A) Linear correlation between COX/nDNA and the mtDNA copy number in six tissues of three individuals (S7/11, S12/12 and S13/12) ($P < 0.0001$). (B) Linear correlation between CS/nDNA and mtDNA copy number in six tissues of three individuals (S7/11, S12/12 and S13/12) ($P < 0.0001$). Correlation was calculated using a bivariate linear fit analysis.

each tissue, expressed either as a percentage of non ribosomal mitochondrial reads on the overall number of mapped reads, or as the sum of RPKM values associated with non ribosomal genes, showed a remarkable correlation with the estimated mtDNA copy number, yielding a Pearson correlation of 0.81, as shown in Fig. 5 (Supplemental Table S10). Similar correlation values (always around 0.8) were obtained by employing different measures for mitochondrial RNA concentration, namely, sum of genes RPKM values including rRNA genes, or fraction of mitochondrial RNA in the RNA-Seq sample, and also when computed against estimated mtDNA abundance derived from qPCR or exome sequencing (data not shown).

4. Discussion

The advent of high-throughput technologies for DNA and RNA sequencing has opened new avenues in biological research. A better understanding of the coordinated expression of the mitochondrial and nuclear genome will be critical for the characterization of novel processes underlying the functioning of eukaryotic cells. In particular, mtDNA copy number is a key functional parameter that varies greatly between different cell or tissues types, both in response to overall metabolic and bioenergetics demands and as a consequence or cause of specific physiological and pathological conditions. Alterations in mtDNA copy number have been related to aging (He et al., 2014), cancer (Zhang et al., 2013), neurodegenerative diseases (Podlesniy et al., 2013), diabetes (Chien et al., 2012) and other mitochondrial-related diseases (Liu et al., 2013).

We previously reported a simple methodology for the reconstruction of complete or nearly complete mitochondrial genomes from off-target reads generated in whole exome sequencing experiments (WES) (Picardi and Pesole, 2012).

Here, using WES data from six different tissues (brain, liver, lung, striated muscle, kidney and heart) from three unrelated healthy Caucasian individuals (males, aged 47–54) in conjunction with qPCR for experimental validation, we show that mtDNA copy number per cell can be reliably estimated from the relative amount of off-target reads of mitochondrial origin detected in WES data. As previously noted (Picardi and Pesole, 2012), it is worth underlining that the fraction of off-target mitochondrial reads is dependent on the exome enrichment protocol used, with the Illumina TruSeq Exome Enrichment Kit providing a higher amount of mitochondrial off-target reads than either Roche Nimblegen and Agilent SureSelect (Picardi and Pesole, 2012). We observed, as expected, strict tissue specificity of mtDNA copy number, reflecting the specific bioenergetics and metabolic demands of different tissues with heart and lung tissues showing the highest and lowest values, respectively. Inter-individual variations for the same tissues

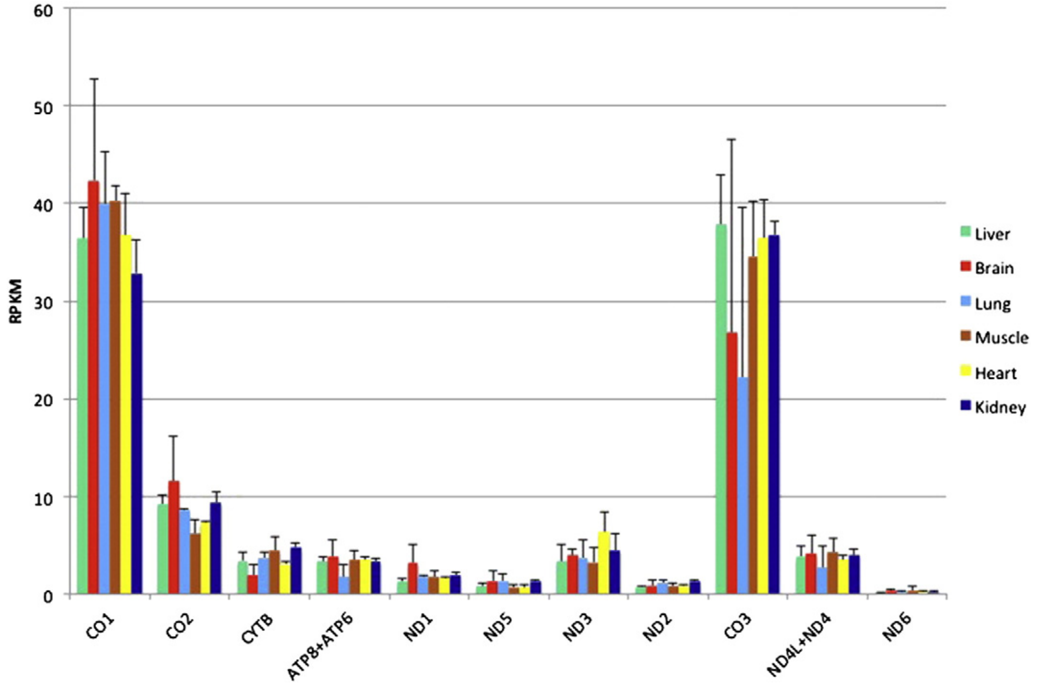


Fig. 4. Relative levels of the mitochondrial mature mRNAs in the six examined tissues expressed as average for the three individuals (S7/11, S12/12 and S13/12). Values represent the ratio between the RPKM (reads per kilobase per million reads) of a specific mRNA and the total RPKM for all mature mRNAs (see Supplementary Table S9) calculated for each given tissue and individual.

were also observed, and are likely related to the individual-specific physiological status and to the cause of death.

Interestingly, we did not detect somatic variations in mtDNA sequences in different tissues from the same individual. Furthermore, the accuracy of sequence data was demonstrated by the overall consistency of gene and transcript sequences obtained by WES and RNA-Seq, respectively. A few putative post-transcriptional changes were detected, most of them in rRNA and tRNA genes, at sites already reported in the literature as susceptible to base-specific post-transcriptional modification.

We also show, by analyzing RNA-Seq data generated from the 18 samples, that a significant correlation exists, between the mtDNA copy number and the expression level of mt protein coding genes. Indeed, the paired-end strand-specific transcriptome sequencing allowed us to obtain reliable estimates of mature and precursor transcripts in sense and antisense orientations. We observed considerable variability in the expression levels of different protein coding transcripts, with CO1 and CO3 significantly more abundant than other genes (Fig. 4). On the other hand, the ND6 gene, the only L-strand gene encoding a polyA- transcript, is expressed at very low levels, totaling on average only 0.22% of all protein coding transcripts. Our data are not in accordance with the expression profile of mitochondrial mRNAs observed in Hela (Piechota et al., 2006) or 143B cells (Mercer et al., 2011). However, it is quite expected that mitochondria from cultured human cells, particularly HeLa cells, do not reflect the transcriptional pattern observed in human tissues.

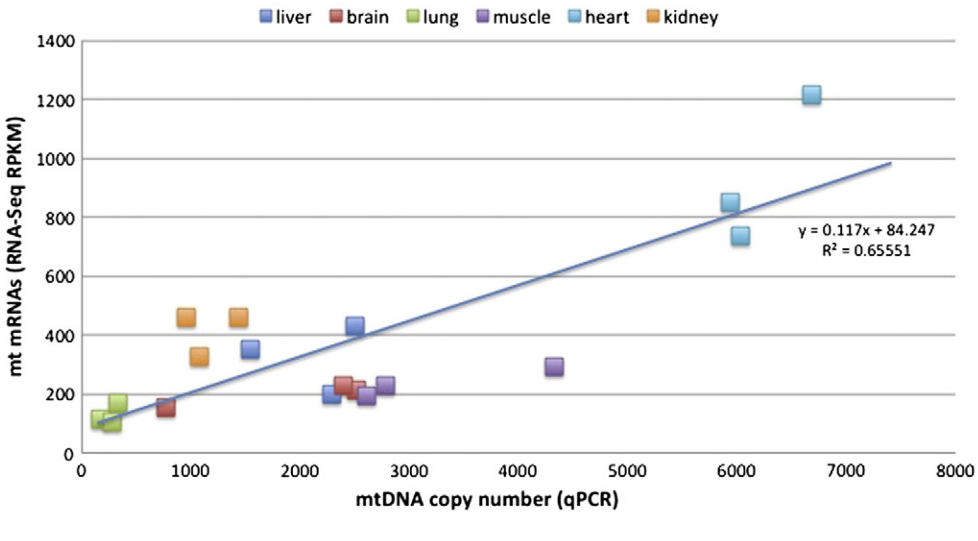


Fig. 5. Linear correlation between estimated mtDNA copy number and relative levels of mt mature transcripts expressed as average of RPKM values in the three individuals (S7/11, S12/12 and S13/12). Correlation was calculated using a bivariate linear fit analysis ($P < 0.001$).

The strand-specific sequencing performed also allowed the detection of stable antisense transcripts of ND5 and ND6 genes (Supplemental Table S9) as already identified in rat (Tullo et al., 1994) and human cell 143B cell line (Mercer et al., 2011). Interestingly, these antisense transcripts have different levels in the different tissues, with the highest levels in heart and kidney.

Moreover, we found a remarkably significant positive correlation between mtDNA copy number and the mitochondrial mass, i.e., the CS/nDNA ratio, as well as the respiratory capacity, i.e., the COX/nDNA ratio, when all the tissues were considered together (Fig. 3), indicating that tissues with the highest mtDNA content i.e., heart, muscle and brain, have also the highest CS and COX activities per cell. Indeed, it is known that different metabolic profiles and variable energetic demands of different tissues are due to the inherently different functions or, in a single tissue, to changes in ATP demand due to physiological or pathological conditions (Kunz, 2003; Leary et al., 1998; Leverve and Fontaine, 2001; Pfeiffer et al., 2001). While our data suggest that mtDNA copy number is a critical parameter in the definition of quantitative aspects of respiratory function, mitochondria also exhibit tissue related variation in functional capacity, protein composition and morphology (Benard et al., 2006; Fernandez-Vizarra et al., 2011; Johnson et al., 2007a, 2007b; Mootha et al., 2003; Pagliarini et al., 2008).

An aspect that must not be neglected—in the light of different metabolic profile of the three individuals—is that cellular adaptation to environmental and physiological constraints necessitates a fine tuning of the control of mitochondrial respiration, in response to changes in energy demand and substrate delivery. Accordingly, different tissues present large differences in the composition of the OXPHOS machinery and the organization of mitochondria (Chan, 2006; Fernandez-Vizarra et al., 2011), which, in addition to mtDNA copy number (Di Mauro and Bonilla, 2004), could influence their physiological activity. Nonetheless, mtDNA copy number remains a good proxy for mitochondrial activity for a sound conceptual reason: the mitochondrial mass or the activity of a respiratory enzyme are being related to the unit of genome that is needed to produce them (Fernandez-Vizarra et al., 2008, 2011) as well as to its relevant nuclear counterparts (Mercer et al., 2011).

In conclusion, the methodology presented here demonstrates the feasibility of large-scale detection of mtDNA copy number in diverse cell-types, tissues and pathological conditions. Given that several thousand exome sequence data sets are available in public repositories from different sources, the approach presented here is expected to generate a wealth of information that may contribute to a better understanding of nucleo-mitochondrion cross-talk and its involvement in health and disease.

Funding

This work was supported by Ministero dell'Istruzione, Università e Ricerca (projects PRIN-2009, Micromap [PON01_02589], Virtualab [PON01_01297]) and by Consiglio Nazionale delle Ricerche (progetto strategico "Medicina personalizzata", progetto strategico "Invecchiamento", progetto bandiera "Epigen").

Data Access

All short read data are available to the users upon request and are being submitted at the dbGAP archive.

Disclosure declaration

The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mito.2014.10.005.

References

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., Howell, N., 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 23, 147.

Bar-Yaacov, D., Avital, G., Levin, L., Richards, A.L., Hachen, N., Rebolledo Jaramillo, B., Nekrutenko, A., Zarivach, R., Mishmar, D., 2013. RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. Genome Res. 23, 1789–1796.

Benard, G., Faustin, B., Passerieux, E., Galinier, A., Rocher, C., Bellance, N., Delage, J.P., Casteilla, L., Letellier, T., Rossignol, R., 2006. Physiological diversity of mitochondrial oxidative phosphorylation. Am. J. Physiol. Cell Physiol. 291, C1172–C1182.

Bobba, A., Amadoro, G., Valenti, D., Corsetti, V., Lassandro, R., Atlante, A., 2013. Mitochondrial respiratory chain Complexes I and IV are impaired by beta-amyloid via direct interaction and through Complex I-dependent ROS production, respectively. Mitochondrion 13, 298–311.

Bogenhagen, D., Clayton, D.A., 1974. The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. J. Biol. Chem. 249, 7991–7995.

Calabrese, F.M., Simone, D., Attimonelli, M., 2012. Primates and mouse NumtS in the UCSC genome browser. BMC Bioinforma. 13 (Suppl. 4), S15.

Capaldi, R.A., 1990. Structure and function of cytochrome *c* oxidase. Annu. Rev. Biochem. 59, 569–596.

Chan, D.C., 2006. Mitochondrial fusion and fission in mammals. Annu. Rev. Cell Dev. Biol. 22, 79–99.

Chien, M.C., Huang, W.T., Wang, P.W., Liou, C.W., Lin, T.K., Hsieh, C.J., Weng, S.W., 2012. Role of mitochondrial DNA variants and copy number in diabetic atherogenesis. Genet. Mol. Res. 11, 3339–3348.

Clay Montier, L.L., Deng, J.J., Bai, Y., 2009. Number matters: control of mammalian mitochondrial DNA copy number. J. Genet. Genomics 36, 125–131.

Di Mauro, S., Bonilla, E., 2004. In: Engel, A., C.F.-A. (Eds.), Myology vol. II. Mc Grav Hill, Philadelphia, pp. 1623–1676.

Draxl, A., Eigentler, A., Gnaiger, E., 2013. PBI-Shredder HRR-Set: preparation of tissue homogenates for diagnosis of mitochondrial respiratory function. Mitochondrial Physiol. Netw. 17 (02), 1–8.

Fernandez-Vizarra, E., Enriquez, J.A., Perez-Martos, A., Montoya, J., Fernandez-Silva, P., 2008. Mitochondrial gene expression is regulated at multiple levels and differentially in the heart and liver by thyroid hormones. Curr. Genet. 54, 13–22.

Fernandez-Vizarra, E., Enriquez, J.A., Perez-Martos, A., Montoya, J., Fernandez-Silva, P., 2011. Tissue-specific differences in mitochondrial activity and biogenesis. Mitochondrion 11, 207–213.

Figueiredo, P.A., Ferreira, R.M., Appell, H.J., Duarte, J.A., 2008. Age-induced morphological, biochemical, and functional alterations in isolated mitochondria from murine skeletal muscle. J. Gerontol. A Biol. Sci. Med. Sci. 63 (4), 350–359.

Goffart, S., von Kleist-Retzow, J.C., Wiesner, R.J., 2004. Regulation of mitochondrial proliferation in the heart: power-plant failure contributes to cardiac failure in hypertrophy. Cardiovasc. Res. 64, 198–207.

He, Y.H., Lu, X., Wu, H., Cai, W.W., Yang, L.Q., Xu, L.Y., Sun, H.P., Kong, Q.P., 2014. Mitochondrial DNA content contributes to healthy aging in Chinese: a study from nonagenarians and centenarians. Neurobiol. Aging 35 (7), 1779.e1–1779.e4.

Johnson, D.T., Harris, R.A., Blair, P.V., Balaban, R.S., 2007a. Functional consequences of mitochondrial proteome heterogeneity. Am. J. Physiol. Cell Physiol. 292, C698–C707.

Johnson, D.T., Harris, R.A., French, S., Blair, P.V., You, J., Bemis, K.G., Wang, M., Balaban, R.S., 2007b. Tissue heterogeneity of the mammalian mitochondrial proteome. Am. J. Physiol. Cell Physiol. 292, C689–C697.

Kirby, D.M., Thorburn, D.R., Turnbull, D.M., Taylor, R.W., 2007. Biochemical assays of respiratory chain complex activity. Methods Cell Biol. 80, 93–119. http://dx.doi.org/10.1016/S0091-679X(06)80004-X.

Kunz, W.S., 2003. Different metabolic properties of mitochondrial oxidative phosphorylation in different cell types—mportant implications for mitochondrial cytopathies. Exp. Physiol. 88, 149–154.

Kunz, W.S., Kudin, A., Vielhaber, S., Elger, C.E., Attardi, G., Villani, G., 2000. Flux control of cytochrome *c* oxidase in human skeletal muscle. J. Biol. Chem. 275, 27741–27745.

Lane, N., Martin, W., 2010. The energetics of genome complexity. Nature 467, 929–934.

Larsen, S., Nielsen, J., Hansen, C.N., Nielsen, L.B., Wibrand, F., Stride, N., Schroder, H.D., Boushel, R., Helge, J.W., Dela, F., Hey-Mogensen, M., 2012. Biomarkers of mitochondrial content in skeletal muscle of healthy young human subjects. J. Physiol. 590, 3349–3360.

Leary, S.C., Battersby, B.J., Moyes, C.D., 1998. Inter-tissue differences in mitochondrial enzyme activity, RNA and DNA in rainbow trout (*Oncorhynchus mykiss*). J. Exp. Biol. 201 (Pt 24), 3377–3384.

Leverve, X.M., Fontaine, E., 2001. Role of substrates in the regulation of mitochondrial function in situ. IUBMB Life 52, 221–229.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

Liu, H., Ma, Y., Fang, F., Zhang, Y., Zou, L., Yang, Y., Zhu, S., Wang, S., Zheng, X., Pei, P., et al., 2013. Wild-type mitochondrial DNA copy number in urinary cells as a useful marker for diagnosing severity of the mitochondrial diseases. PLoS One 8, e67146.

Mazat, J.P., Rossignol, R., Malgat, M., Rocher, C., Faustin, B., Letellier, T., 2001. What do mitochondrial diseases teach us about normal mitochondrial functions…that we already knew: threshold expression of mitochondrial defects. Biochim. Biophys. Acta 1504, 20–30.

Mercer, T.R., Neph, S., Dinger, M.E., Crawford, J., Smith, M.A., Shearwood, A.M., Haugen, E., Bracken, C.P., Rackham, O., Stamatoyannopoulos, J.A., Filipovska, A., Mattick, J.S., 2011. The human mitochondrial transcriptome. Cell 146, 645–658.

Mootha, V.K., Bunkenborg, J., Olsen, J.V., Hjerrild, M., Wisniewski, J.R., Stahl, E., Bolouri, M.S., Ray, H.N., Sihag, S., Kamal, M., Patterson, N., Lander, E.S., Mann, M., 2003. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. Cell 115, 629–640.

Moraes, C.T., 2001. What regulates mitochondrial DNA copy number in animal cells? Trends Genet. 17, 199–205.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628.

Nardelli, M., Tommasi, S., D'Erchia, A.M., Tanzariello, F., Tullo, A., Primavera, A.T., De Lena, M., Sbisa, E., Saccone, C., 1994. Detection of novel transcripts in the human mitochondrial DNA region coding for ATPase8-ATPase6 subunits. FEBS Lett. 344, 10–14.

Pagliarini, D.J., Calvo, S.E., Chang, B., Sheth, S.A., Vafai, S.B., Ong, S.E., Walford, G.A., Sugiana, C., Boneh, A., Chen, W.K., et al., 2008. A mitochondrial protein compendium elucidates complex I disease biology. Cell 134, 112–123.

Pesole, G., Allen, J.F., Lane, N., Martin, W., Rand, D.M., Schatz, G., Saccone, C., 2012. The neglected genome. EMBO Rep. 13, 473–474.

Pfeiffer, T., Schuster, S., Bonhoeffer, S., 2001. Cooperation and competition in the evolution of ATP-producing pathways. Science 292, 504–507.

Picardi, E., Pesole, G., 2012. Mitochondrial genomes gleaned from human whole-exome sequencing. Nat. Methods 9, 523–524.

Picardi, E., Pesole, G., 2013. REDItools: high-throughput RNA editing detection made easy. Bioinformatics 29, 1813–1814.

Piechota, J., Tomecki, R., Gewartowski, K., Szczesny, R., Dmochowska, A., Kudla, M., Dybczynska, L., Stepien, P.P., Bartnik, E., 2006. Differential stability of mitochondrial mRNA in HeLa cells. Acta Biochim. Pol. 53, 157–168.

Podlesniy, P., Figueiro-Silva, J., Llado, A., Antonell, A., Sanchez-Valle, R., Alcolea, D., Lleo, A., Molinuevo, J.L., Serra, N., Trullas, R., 2013. Low cerebrospinal fluid concentration of mitochondrial DNA in preclinical Alzheimer disease. Ann. Neurol. 74, 655–668.

Rackham, O., Shearwood, A.M., Mercer, T.R., Davies, S.M., Mattick, J.S., Filipovska, A., 2011. Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. RNA 17, 2085–2093.

Rafelski, S.M., 2013. Mitochondrial network morphology: building an integrative, geometrical view. BMC Biol. 11, 71.

Robinson Jr., J.B., Srere, P.A., 1985. Organization of Krebs tricarboxylic acid cycle enzymes in mitochondria. J. Biol. Chem. 260, 10800–10805.

Ruiz-Pesini, E., Lott, M.T., Procaccio, V., Poole, J.C., Brandon, M.C., Mishmar, D., Yi, C., Kreuziger, J., Baldi, P., Wallace, D.C., 2007. An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. 35, D823–D828.

Sarnat, H.B., Marin-Garcia, J., 2005. Pathology of mitochondrial encephalomyopathies. The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques 32 (2), 152–166.

Shmookler Reis, R.J., Goldstein, S., 1983. Mitochondrial DNA in mortal and immortal human cells. Genome number, integrity, and methylation. J. Biol. Chem. 258, 9078–9085.

Tullo, A., Tanzariello, F., D'Erchia, A.M., Nardelli, M., Papeo, P.A., Sbisa, E., Saccone, C., 1994. Transcription of rat mitochondrial NADH-dehydrogenase subunits. Presence of antisense and precursor RNA species. FEBS Lett. 354, 30–36.

Van den Bogert, C., De Vries, H., Holtrop, M., Muus, P., Dekker, H.L., Van Galen, M.J., Bolhuis, P.A., Taanman, J.W., 1993. Regulation of the expression of mitochondrial proteins: relationship between mtDNA copy number and cytochrome-c oxidase activity in human cells and tissues. Biochim. Biophys. Acta 1144, 177–183.

Venegas, V., Wang, J., Dimmock, D., Wong, L.J., 2011. Real-time quantitative PCR analysis of mitochondrial DNA content. In: Haines, Jonathan L., et al. (Eds.), Current protocols in human genetics/editorial board, p. 17 (Chapter 19, Unit 19).

Villani, G., Attardi, G., 1997. In vivo control of respiration by cytochrome *c* oxidase in wild-type and mitochondrial DNA mutation-carrying human cells. Proc. Natl. Acad. Sci. U. S. A. 94, 1166–1171.

Villani, G., Attardi, G., 2000. In vivo control of respiration by cytochrome *c* oxidase in human cells. Free Radic. Biol. Med. 29, 202–210.

Villani, G., Greco, M., Papa, S., Attardi, G., 1998. Low reserve of cytochrome *c* oxidase capacity in vivo in the respiratory chain of a variety of human cell types. J. Biol. Chem. 273, 31829–31836.

Waddell, W.J., 1956. A simple ultraviolet spectrophotometric method for the determination of protein. J. Lab. Clin. Med. 48, 311–314.

Wu, T.D., Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26, 873–881.

Zhang, G., Qu, Y., Dang, S., Yang, Q., Shi, B., Hou, P., 2013. Variable copy number of mitochondrial DNA (mtDNA) predicts worse prognosis in advanced gastric cancer patients. Diagn. Pathol. 8, 173.

Acknowledgements

I owe my gratitude to my supervisor Prof. Giulio Pavesi whose teaching and support in the last three years have been invaluable.
His brilliant advices on both programming and biology have guided me in the complex field of bioinformatics, and I will be forever indebted to him for all the time and patience he invested in my personal formation.

It was an honor to be (a small) part of Prof. Graziano Pesole projects, which allowed me to complete my doctoral studies.

I also would like to thank my two awesome colleagues, Federico Zambelli and Matteo Chiara, whose clever tips on programming and bionformatics, conversations on geek stuff and witty sense humour accompanied me in this long journey. Thank you for the amazing moments in the last 3 years and… for all the fish! [cit.]

I owe my deepest gratitude to my parents, who supported me until now, in both the good and (especially) the bad times. Their love and caring have guided me through many difficulties, not just in my PhD years but in my whole life. One couldn't dream of better parents.

I also would like to thank my many friends, especially Marco and Simone for all the great time we had together, without you my life wouldn't be complete (and definitely much much sadder and lonelier). Your support and your ability to make me forget all my problems with a few laughs are pure gold. Thank you Alessandro for your support and meal company, I wish the best in life to you and to all my friends in Lodi. Finally, I would like to thank Marco for the neverending gaming sessions in the weekends and for all the awesome people I got to know because of you.