Methods 48 (2009) 249–257

Contents lists available at ScienceDirect

Methods

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

Review Article

RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing

Brian T. Wilhelm [a,b,*], Josette-Renée Landry [b]

[a] Laboratory of Molecular Genetics of Stem Cells, C.P. 6128 Succursale Centre-Ville, Montréal, Que., Canada H3C 3J7
[b] Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, C.P. 6128 Succursale Centre-Ville, Montréal, Que., Canada H3C 3J7

ARTICLE INFO

Article history:
Accepted 17 March 2009
Available online 29 March 2009

Keywords:
Massively parallel sequencing
RNA-seq
Transcriptomics
Next generation sequencing

ABSTRACT

The ability to quantitatively survey the global behavior of transcriptomes has been a key milestone in the field of systems biology, enabled by the advent of DNA microarrays. While this approach has literally transformed our vision and approach to cellular physiology, microarray technology has always been limited by the requirement to decide, *a priori*, what regions of the genome to examine. While very high density tiling arrays have reduced this limitation for simpler organisms, it remains an obstacle for larger, more complex, eukaryotic genomes.

The recent development of "next-generation" massively parallel sequencing (MPS) technologies by companies such as Roche (454 GS FLX), Illumina (Genome Analyzer II), and ABI (AB SOLiD) has completely transformed the way in which quantitative transcriptomics can be done. These new technologies have reduced both the cost-per-reaction and time required by orders of magnitude, making the use of sequencing a cost-effective option for many experimental approaches. One such method that has recently been developed uses MPS technology to directly survey the RNA content of cells, without requiring any of the traditional cloning associated with EST sequencing. This approach, called "RNA-seq", can generate quantitative expression scores that are comparable to microarrays, with the added benefit that the entire transcriptome is surveyed without the requirement of *a priori* knowledge of transcribed regions. The important advantage of this technique is that not only can quantitative expression measures be made, but transcript structures including alternatively spliced transcript isoforms, can also be identified. This article discusses the experimental approach for both sample preparation and data analysis for the technique of RNA-seq.

1. Introduction

Since their development little more than a decade ago, DNA microarrays have provided scientists with the capacity to simultaneously investigate thousands of features in a single experiment. This capability has been exploited not only to monitor the steady state expression of genes, but also to locate regions of copy number changes in cancers (array-based CGH) [1], to map the genome-wide binding sites of DNA interacting proteins (ChIP-on-chip) [2,3] and to survey long range DNA interactions (4C) [4]. The overwhelming wealth of knowledge generated by microarrays has created entirely new fields of research and, as the underlying technology became broadly adopted, microarrays forever changed the way in which high throughput science is done.

Equally revolutionary technologies are currently emerging in the form of new methods of sequencing, termed massively parallel sequencing (MPS, also called next-generation/ultra high throughput sequencing). Three brands of machines based on these new technologies currently dominate the field (454 GS FLX (Roche), Genome Analyzer II (Illumina) and SOLiD (Applied Biosystems)), with others on the horizon with potentially even higher outputs (Pacific Biosciences, Helicos). While the individual approaches vary considerably in their technical details, the essence of these systems is the miniaturization of individual sequencing reactions. Each of these miniaturized reactions is "seeded" with DNA molecules, at limiting dilutions, such that there is a single DNA molecule in each, which is first amplified and then sequenced. The physical design of these instruments allows for an optimal spatial arrangement of each reaction, enabling an efficient readout by laser scanning (or other methods), for millions of individual sequencing reactions to be put onto a standard glass slide. While the immense volume of data generated is attractive, it is arguably the elimination of the cloning step of the DNA fragment to be sequenced that is the greatest benefit of these new technologies. All of the current methods allow the direct use of small fragments of DNA without a requirement for insertion into a plasmid or other vector, thereby removing a costly and time consuming step required for traditional Sanger sequencing.

* Corresponding author. Address: Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, C.P. 6128 Succursale Centre-Ville, Montréal, Que., Canada H3C 3J7. Fax: +1 514 343 7780.
E-mail address: Brian.wilhelm@umontreal.ca (B.T. Wilhelm).

1046-2023/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2009.03.016

Predictably, MPS technology has already been used to perform *de novo* genome sequencing [5] and to complete whole genome scans of protein–DNA interaction sites [6,7] (through sequencing ChIP material instead of hybridizing it; ChIP-seq vs. ChIP-chip) and a variety of other applications [8,9]. More recently, an increasing number of studies (a subset of which are shown in Table 1) have demonstrated the potential for the sequencing of total cDNA in order to observe the complete transcriptome. These studies have clearly demonstrated the advantages of this approach: in a single RNA-seq experiment, one can derive not only an accurate, quantitative measure of individual gene expression (as with a standard expression microarray), but also discover novel transcribed regions in an unbiased manner (as with a whole genome tiling approach). In addition, this methodology enables a global survey of the usage of the alternative splice sites [10,11] (similar to a custom designed splicing array).

Based on the demonstrated power of the RNA-seq approach it is clear that, at least for comprehensive studies in higher eukaryotes,

Table 1
Selection of RNA-seq studies to date.

Organism	Platform	Total reads (10^6)	Mapped reads (10^6)	Main study focus	Growth conditions/cell types	Priming/preparation method	Ref.
Fission yeast	Illumina	23, 99	104	Transcriptome characterization, transcriptome changes related to differentiation	Rich media, five stages of meiotic differentiation	PolyA selection, oligo dT primed	[16]
Budding yeast	Illumina	30	16	Transcriptome characterization	Rich media	PolyA selection, random primed or oligo dT primed	[24]
Mouse	Illumina	78, 71, 67	52, 42, 43	Transcriptome changes related to differentiation	Adult mouse brain, liver and skeletal muscle tissues	PolyA selection, random primed	[15]
Mouse	ABI SOLiD	4.9, 5.0	2.9, 2.8	Transcriptome changes related to differentiation	Undifferentiated mouse embryonic stem cells (ESCs) and embryoid bodies	rRNA depletion, PolyA selection, random primed and tagged for direction	[14]
Human cell line	Illumina	8.6, 7.7	6.4, 5.4	Alternative splicing diversity	Human embryonic kidney and B cell line	PolyA selection, random primed	[11]
Human	454	0.18	0.14	Transcriptome changes related to differentiation	Prostate cancer cell line treated with androgens	Not provided	[28]
Arabidopsis	454	0.54	0.48	Transcriptome characterization	Aerial tissues of 8d old seedlings	PolyA selection, oligo dT primed	[29]
Human	454	2.4–2.9	1.6–2.4	Disease characterization (mutation discovery, expression differences)	Malignant pleural mesothelioma (4), adenocarcinoma (1) and normal lung (1)	PolyA selection, oligo dT primed	[30]
Arabidopsis	Illumina	83, 88, 92, 106 (mRNA) 8.7, 8.8, 12.2, 7.5 (smRNA)	56, 47, 46, 49 (mRNA) 2.8, 2.5, 2.1, 3.5 (smRNA)	Transcriptome characterization, integrating RNA, small RNA, and CpG methylation data	Immature floral tissue (mRNA, smRNA from wt, and met1, ddc, rdd mutants)	rRNA depletion and directionally ligated for strand specificity (mRNA), RNA linker ligation, RT-PCR (smRNA)	[18]
Locust	Illumina	1.5, 1.9 (smRNA)	Not mapped	Small RNA transcriptome changes related to differentiation	Gregarious locus, solitary locus	RNA linker ligation, RT-PCR	[31]
Human	Illumina	23	10.8	Global promoter proximal transcription	Lung fibroblast (global run-on)	BrU incorporation, base hydrolysis, immuno-purification, sequential RNA linker ligation, RT-PCR	[32]
Human	Illumina	28.6	18.9	Transcriptome characterization	HeLa S3	(1) rRNA depletion,+/− PolyA selection, random primed. (2) Oligo dT priming, dscDNA sonication, linker ligation	[33]
Mouse	Illumina	3.0	2.3	Transcriptome characterization	ES cells	PolyA selection, oligo dT primed	[34]
Grape vine	Illumina	173	138	Transcriptome characterization	Leaf, root, stem, callus	PolyA selection, random primed	[35]
Human	Illumina	11.7, 8.5	3.4, 2.8	Transcriptome changes related to differentiation	LNCaP, +/− DHT	PolyA selection, double random priming with modified (P1 biotinylated) sequencing primers	[36]
Human	Illumina	6.1, 6.0	0.77, 0.72	Small RNA transcriptome changes related to differentiation	Undifferentiated human embryonic stem cells (ESCs) and embryoid bodies	Size selection, sequential RNA linker ligation, RT-PCR	[37]
Sea anemone, Sponge, Placozoan, choanoflagellate	454 Illumina	Not provided	7.8 for all species/ conditions	Small RNAs, small RNA transcriptome changes during evolution	Mixed developmental stages (Nematostella, Trichoplax), separate adult/embryonic stages (Amphimedon)+/− periodate treated (Nematpstella, Amphimedon)	Size selection, sequential RNA linker ligation, RT-PCR	[38]
Human	454	0.55, 0.24, 0.83 (454), 67, 76, 57, 72, 14, 35, 9 (Illumina)	38, 40, 35, 45, 10, 16, 6 (Illumina)	Gene fusions in cancer	VCaP, LNCaP & RWPE (454), K562,VCaP, LNCaP, RWPE, VCaP-Met, Met3, Met4 (Illumina)	PolyA selection, RT primers not specified	[39]
Human	Illumina/ Solexa	75, 75	29, 30	Technical assessment of ability to measure transcriptome changes related to differentiation	Liver and kidney samples	PolyA selection, random primed	[40]

where surveys of differential splicing activity, antisense transcription, and discovery of novel regions of transcription are desired, high throughput sequencing of RNA will quickly supersede microarray-based methods. A comparison of microarray and sequencing based techniques for measuring gene expression levels is shown in Table 2. In addition to the inclusive features of an RNA-seq experiment described above, the two principle differences between sequencing and microarray approaches are in the resolution of the output, and in the dynamic range of changes that can be observed. While very high resolution tiling arrays are available for some simple organisms, most array designs for higher eukaryotes have resolutions $\geqslant$35 bp, while the output from RNA-seq experiments is already at the theoretical maximum of base-pair resolution. With regards to the dynamic range of expression values which can be obtained, the limits of most microarray scanners means that only several orders of magnitude of expression signals can be measured. In RNA-seq experiments, the limits of the dynamic range measured are only determined by the amount of sequencing obtained. This means that through the continued sequencing of a given library, it should be possible to eventually measure the expression of every transcript present and so the "dynamic range" only represents the actual biological diversity of the transcriptome.

While this unprecedented level of sensitivity brings with it the power to make many novel biological observations, it also carries the price of vastly increased bioinformatics challenges in dealing with the massive data files and extracting biologically relevant data. As noted in Table 2, the raw image files from one run of some next generation sequencers can require terabytes of storage (version 3 of the ABI SOLiD ships with 12 Tb of "on machine" storage for instance), meaning that simply moving the data off the machine can represent a technical challenge for the computer networks of many research centers. More over, even when the data is moved off the machine for subsequent processing, even a high-end desktop computer will be hopelessly outmatched by the volume of data from a single run, in terms of being able to carry out all aspects of the analysis in a reasonable amount of time. As a result, although non-trivial to establish, the use of a small cluster of computers is extremely beneficial, in order to remove computational bottlenecks, as discussed later. One final issue related to dealing with the data from an RNA-seq experiments is the software required to perform downstream analysis. Given how novel the next-generation sequencing technology is, it is not surprising that there are no "box standard" software packages available for end-users, hence software is often developed on an *ad hoc* basis. While some software packages are beginning to appear that enable some general aspects of RNA-seq analysis to be performed, these are still generally only useful for labs with fairly strong pre-existing bioinformatic capabilities. While the computational challenges mentioned above are not often given much space in discussions of next-generation sequencers and their potential, they do actually present immediate (and rapidly increasing) road blocks to the use of this new technology.

Table 2
Comparison of current methods for surveying transcriptome.

Criterion	Expression arrays	Tiling arrays	RNA-Seq
Resolution of data	N/A	Dependent on genome size but $\geqslant$35 bp for human/mouse	1 bp, at sufficient sequencing depth
Cost per sample (excluding equipment)	Low	Low–high, depending on arrays needed to cover genome	High
Linear dynamic range of expression values	<4 orders of magnitude	<2 orders of magnitude	Limited only by sequencing depth and biological expression levels
Sensitivity (Signal:Noise)	Moderate	Low	High
Discovery of novel transcribed regions	No	Yes	Yes
Monitor splice site usage	No	Limited	Yes
Identification alternative promoters/ UTRs	No	Yes	Yes
Detection of antisense transcripts	Not standard	Not standard	Requires strand specific preparation
Detection of SNPs, mutations, allelic differences	Limited	Limited	Yes
Size of raw data files per experiment	0.01–0.05 Gb	0.1–1 Gb	1–15 Tb
Downstream Bioinformatic requirements	Low	High	Very high

2. Sample preparation

Given the variety of current technical approaches (many of which may be obsolete before this article is published), a precise step-by-step protocol would not be particularly practical for a methodology paper. Instead, this article will focus on the key elements of the procedure which are common to all technologies, and discuss the factors which should be considered when planning such experiments. A general overview of the work flow is shown in Fig. 1.

2.1. Amount requirements

Because the RNA-seq approach is entirely based on the general principles of DNA sequencing, the methodology should be applicable to any organism, subject to the availability of a sufficient amount of RNA. It is worth noting that while published information on the performance of these technologies in high/low GC content genomes is scarce, anecdotally, they do not appear to show any significant bias across a fairly wide spread of GC content (30–70%), suggesting that RNA from most organisms would be suitable. In the event that RNA is limiting, approaches to amplify small quantities of RNA exist [12,13], which should enable the use of such samples, with the caveat that, depending on the extent of the amplification, biases could be introduced in the sample. As discussed below, the amount of RNA required is dependent on both the sequencing technology and the method of priming used. Most studies to date have utilized PolyA enrichment steps to selectively remove ribosomal RNA [14–16] (see below) such that 100–200 ng of PolyA $^+$ enriched RNA (derived from 100 μg of total RNA) is used for double-stranded cDNA synthesis prior to sequencing.

2.2. rRNA removal

One of the principal technical hurdles to overcome with RNA-seq is the fact that the vast majority of RNA (>90%) present in cells consists of ribosomal RNA (rRNA). As such, the bulk of the total RNA is not informative as to the true diversity of the transcriptome present in the remaining RNA. In order to avoid wasting effort in re-sequencing the same ribosomal RNA millions of times, several techniques exist to focus the sequencing effort on the non-ribosomal portion.

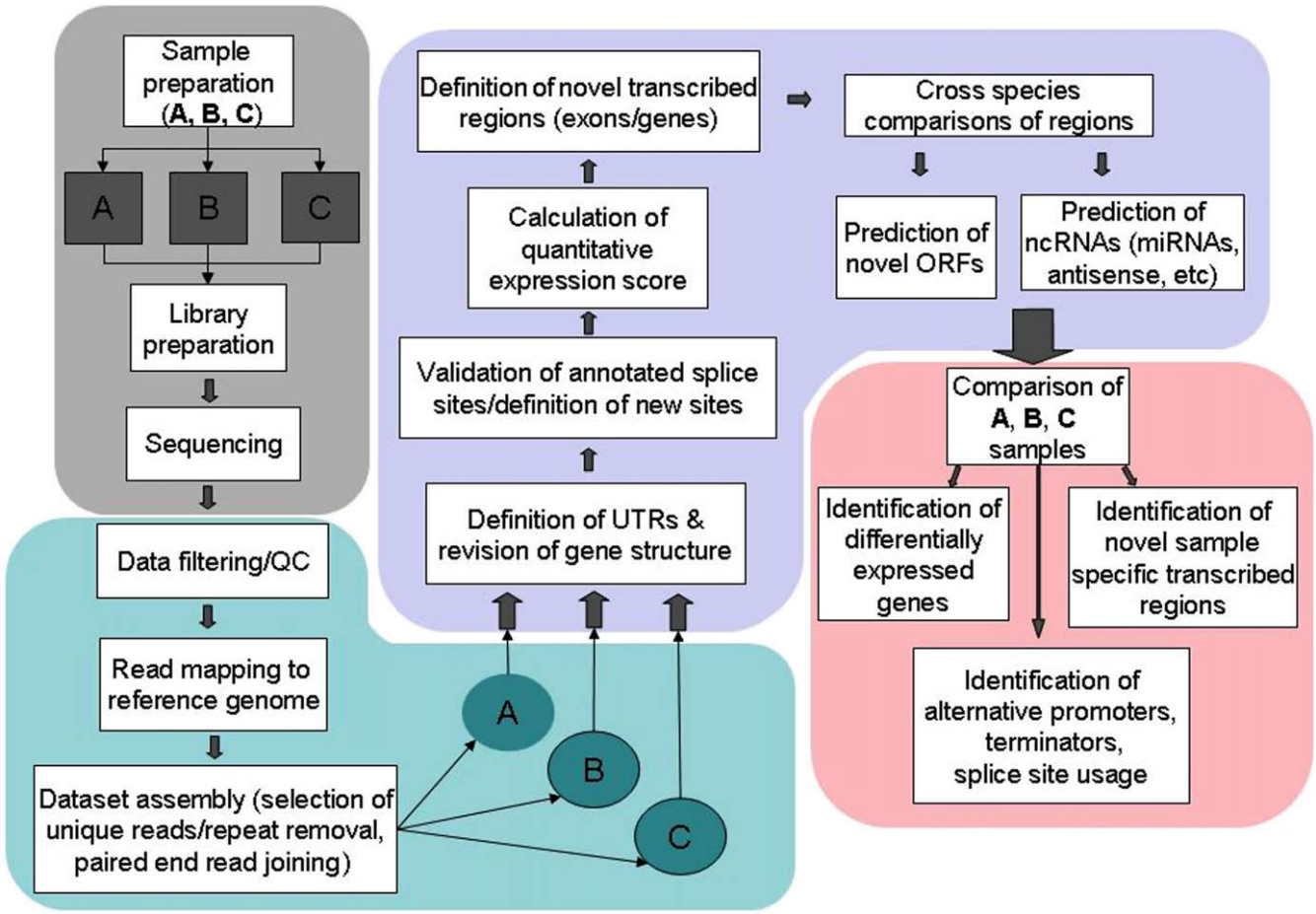

Fig. 1. Example of RNA-seq work flow. A typical analysis stream for three theoretical biological samples (A, B and C) is shown with the various sections color coded as wet-lab work (grey), creation of the filtered data sets (cyan), sample (tissue/developmental stage/growth condition) specific data analysis (lilac) and cross sample analysis (pink). Much of the later analysis (lilac/pink) will be highly dependent on the experimental aims of the study and as such, only a small fraction of the possible analyses pathways are shown.

One option is to selectively enrich for mRNA present in the total RNA. This can be accomplished using various commercial kits which either selectively remove rRNA of selectively enrich for mRNA. The rRNA depletion kits (e.g. RiboMinus (Invitrogen)) use antisense versions of the ribosomal transcripts to be removed which are conjugated to either biotinylated or magnetic beads. In the case of polyA enrichment, similar beads that have long oligo dT stretches (e.g. OligoTex (Qiagen)), which bind the polyA tails present on most mRNA molecules, can selectively enrich for non-ribosomal transcripts. As a last alternative, enzymes which are capable of selectively degrading uncapped (rRNA) are also commercially available, although no studies published have yet used this method.

2.3. Priming

Following enrichment, the resulting mRNA must be primed for the reverse transcription reaction using either random primers or oligo dT primers. The advantage of using oligo dT (with or without prior mRNA enrichment) is that the majority of cDNA produced should be polyadenylated mRNA, hence more of the sequence obtained should be informative (non-ribosomal). The significant disadvantage of the use of oligo dT primers is that the reverse transcriptase enzyme will fall off of the template at a characteristic rate, resulting in a bias towards the 3′ end of transcripts. For long mRNAs, this bias can be pronounced, resulting in an underrepresentation (or worse, absence) of the extreme 5′ end of the transcript in the data. The use of random primers in general would therefore be the preferred method to avoid this problem and to allow a better representation of the 5′ end of long ORFs. However, when oligo dT primers are used for priming, the slope which is formed by the diminishing frequency of reads towards the 5′ end of the ORF can, in some cases, be useful for determining the strand of origin for novel transcripts if strand information has not been retained as described below.

2.4. Maintaining strand specific information

An additional consideration in the process of creating the double-stranded cDNA for sequencing is to maintain strand specific information for the RNA. The importance of this consideration will obviously vary depending on the organism being studied, but in more complex genomes (such as mouse and human) where there is clear evidence for wide spread antisense transcription [17], strand specific information should be considered a clear requisite for comprehensive RNA-seq studies. To date, few papers have demonstrated a feasible methodology for maintaining strand specific RNA information when creating ds cDNA [14,18]. These methods involved the fragmentation of the enriched mRNA through the use of either metal hydrolysis or heat followed by ligation of RNA linkers. In the case of Cloonan et al., the synthesis of the first cDNA strand from the fragmented RNA is performed using a tagged

random hexamer and a modified MMLV Reverse Transcriptase, which allows additional nucleotides to be added to the termini. After the addition of C nucleotides to the 3′ ends of the first cDNA strand product, another tagged poly G track primer is used to prime the second strand synthesis, resulting in (PCR amplifiable) double-stranded cDNA molecules with tags that mark the original 5′ and 3′ ends. A similar approach was used by Lister et al. although the RNA linkers ligated onto the fragment RNA at either end were the Illumina 3′ & 5′ RNA oligonucleotide adapter sequences, which allowed for RT-PCR amplification (20 cycles) directly and facilitated later processing.

While the use of such end tagging approaches does carry the disadvantage of increased RNA handling and the potential loss of a small amount of "tag" sequence (i.e. non-templated C's) from the short cDNA sequence read (~45 bp), the scientific importance of strand specificity and continual increases in individual read lengths, through improved reaction chemistry, more than balance these concerns.

3. Sequencing

The methodology presented below is based on the Illumina platform and, as such, is not applicable to other next generation sequencing machines. However, all such platforms share the same basic principle: the isolation and attachment to a solid matrix of a single DNA fragment through limiting dilution, followed by amplification of this single molecule either through a specialized emulsion PCR (EM-PCR; SOLiD/454) or a linker based bridging reaction (Illumina). These larger, discreet populations of identical molecules can then be sequenced in parallel, either through the measurement of the incorporation of fluorescent nucleotides (Illumina) or short fluorescent linkers (SOLiD), or through the release of by-products from incorporation of normal nucleotides (454). A detailed comparison of the technical differences (including discussion of some potentially proprietary information) in these rapidly evolving approaches (or others nearing release) is beyond the scope of this article. Nevertheless, a general example of sample preparation has been provided to give readers some idea of what is involved in this stage of the sample processing.

4. Overview of Illumina Genome Analyzer II protocol

(1) Fragmentation of the cDNA. The Illumina recommended method for this step is nebulization. However, in principle, other methods such as controlled DNase digestion should work as well.
(2) Purification of the fragmentation products using Qiagen QIAquick PCR kit or equivalent.
(3) End repair of cDNA fragments. The nebulization (or other fragmentation methods) generates double-stranded cDNA fragments with a mix of blunt-ends as well as 3′/5′ overhangs. The 3′>5′ exonuclease activity of Klenow polymerase and T4 DNA polymerase is used to blunt the ends of all of the fragments.
(4) Tailing of cDNA fragments. Klenow (exo-) is used with dATP in order to add a 3′ adenine overhang to the blunt-end double-stranded cDNA fragments.
(5) Adaptor ligation. Tailed cDNA fragments are ligated to a mix of two adaptors, that will permit the non-specific amplification of cDNA fragments from step 6.
(6) Size-based purification of ligation products. Ligation products are separated on a 2% TAE (Tris–acetate–EDTA)–agarose gel, and a specific region of the gel is excised according to the desired size range for the insert (generally 120–170 nt in size). The cDNA contained in the gel fragment is purified using a gel extraction kit.
(7) PCR of ligation products. Purified cDNA fragments from step 6 are used for seventeen rounds of PCR amplification with primers complementary to the previously ligated adaptors, and compatible to oligonucleotides attached to the Illumina FlowCell.
(8) Purification and quantification. Following PCR amplification of ligation products, the resulting DNA is purified with a QIAquick PCR kit (Qiagen) or equivalent, and concentration is measured with a Nanodrop. The cDNA is subsequently diluted to a working concentration of 10 nM in TE.
(9) Sequencing of fragments. Purified cDNA fragments are loaded onto Illumina flow cells, keeping the 8th flow cell reserved for a recurrent internal standard to control for sequencing efficiency. The instrument is typically run for 35–45 cycles of chemistry (to allow for the incorporation of 35–45 bp, although newer versions of the Genome Analyzer are capable of extending this further).

5. Data analysis

5.1. Data filtering

As part of the results from any sequencing run, there will inevitably be a certain percentage of reads which cannot be mapped (which may be contaminants), or which are of dubious quality (large numbers of Ns called in the sequence). The need to remove these poor quality reads is somewhat arbitrary in that low quality reads are much less likely to be matched to the reference genome, and in general, the percentage of such poor quality reads is relatively low. Nevertheless, the removal of such reads is not difficult and will accelerate subsequent downstream analysis. Reads which contain numerous interspersed Ns in their sequence, or short reads (<~17 bp) lacking Ns which would in any case be too short to be effectively matched back to a reference genome, are unlikely to be informative for any application. A series of simple Perl scripts can be used to remove such low quality sequence reads.

5.2. Read mapping

Once the cDNA sequence reads have been filtered to remove aberrant reads, the next challenge is to match the sequences back to the reference genome. While this task is trivial on an individual basis, it is the sheer volume provided by this methodology that presents a challenge in the context of next-generation sequencing technologies. To illustrate the extent of the problem, it would require either 43 or 6 h, using established alignment programs such as BLAST and BLAT respectively, to map 10 million 32 bp reads, back to a reference genome [19]. Since some next-generation sequencers are capable of generating well over 20 times this number of reads, the absolute requirement for new bioinformatic tools is self-evident. Novel alignment programs, both published (SOAP, MAQ, SSAHA2 [19–21]) and unpublished (ELAND, BOWTIE) are available and are specifically adapted for high volumes of short sequence reads, allowing an enormous advantages in terms of alignment efficiency (1–8 min to map ~10 million reads). Given the nature of the computational task and the structure of the data programs such as BLAT can still be used effectively by breaking down the total number of reads to be mapped into arbitrarily smaller "chunks", if large amounts of distributed computing power is readily available. These subsets of sequence reads can afterwards be sent separately to individual nodes of a compute farm. The final mapping results can then be recompiled from the results of the individual BLAT output files. Despite the technical feasibility of this

solution, the general rarity of such compute farms reduces the applicability of this approach. As a result, the main bioinformatic efforts have been towards the development of more efficient alignment programs which will be capable of running on individual workstations, although this remains a challenging task.

Regardless of the software used, the issue of how to deal with repetitive sequences, especially those that match perfectly to more than one location in the genome, remains. For complex organisms (human and mouse) where repetitive sequences represent nearly 50% of the genome [22,23], the impact of the inclusion of such reads will be dependent on the size of the transcript, its expression level, the distribution of the sequence reads across the transcript and the frequency of the expressed version of the repeated sequence. In many cases, the effect may be well below the noise levels that are present in alternative technologies such as microarrays. However, with so many parameters affecting the data, the simplest approach to repetitive sequences in transcripts, may simply be to remove them from the results unless they can be matched unambiguously. The number of reads which cannot be matched unambiguously can also be substantially reduced through the use of paired-end reads (or mate-pair reads for the ABI SOLiD).

Currently, all of the next generation sequencing technologies are capable for generating data from "paired-end" or "mate-pair" fragments. This approach involves sequencing both ends of a single molecule of an approximately known size (based on the fragment excised in step 6 in the protocol above), thereby creating a larger (~120 bp) pseudo-read with 35–45 bp of known sequence at either end. As with conventional whole genome shotgun sequencing, the sequence information of either end of an approximately known size is extremely useful for mapping reads. When one of the paired reads maps to a highly repetitive element in the genome but the second does not, it allows both reads to be mapped to the reference genome unambiguously. This is accomplished by first matching the first non-repeat read uniquely to a genomic position and then looking within a size window, based on the known size range of the library fragments, for a match for the second read. The usefulness of this approach was demonstrated to improve read matching from 85% (single reads) to 93% (paired reads) [20], allowing a significant improvement in genome coverage, articulacy in repeat regions. RNA-seq analysis software or read mapping programs such as SSAHA2, ERANGE3.0 and MAQ all support the use of paired-end reads.

6. Expression scoring and representation

In order to derive expression scores for annotated elements (such as exons) within a genome, a method must be used to convert RNA-seq reads into a quantitative value for each element. The simplest approach to this problem is to simply sum the number of reads which fall within the co-ordinates of each element (either exon or gene), and then normalize for the length of the element (Fig. 2). An alternative approach is to calculate a sequence score for each nucleotide in the genome based on the number of reads which cover each base position, and again normalizing for element lengths. All methods generate fairly similar results, depending on how the sample is prepared, one methodology may be more suitable (i.e. oligo dT priming may produce poor coverage of the 5′ ends of longer expressed genes in humans, hence counting reads/scores at the 3′ region of all genes may overcome this problem). In addition to normalizing for the feature length, when comparing different RNA-seq data from different conditions or tissues, it is also necessary to normalize for the total amount of sequence obtained per condition. This can be done simply by dividing the scores calculated above for all features by the ratio of sequence depth of one "reference condition", or total sequence for each condition (in bp), or some other equivalent transformation. This avoids the possibility that genes will appear to be differentially expressed simply as a result of the presence of more sequence in one condition as compared to another.

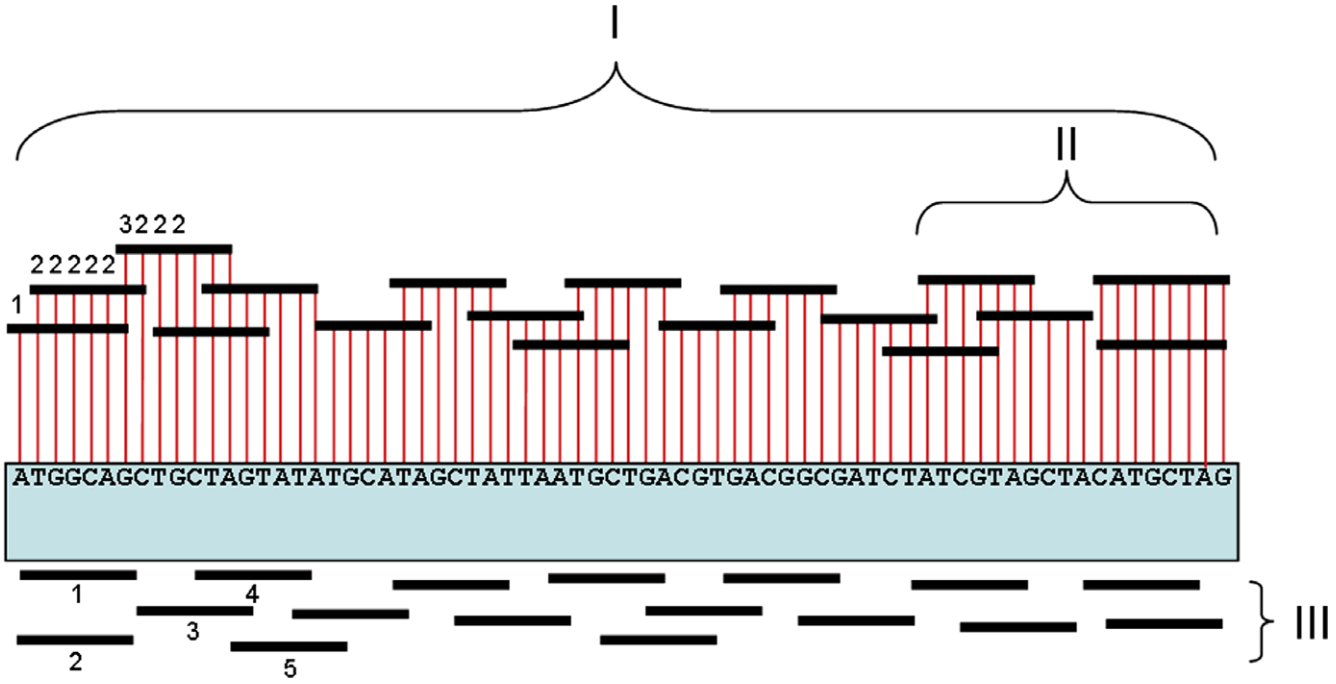

Fig. 2. Different methods of quantifying sequence expression scores. This figure shows three possible methodologies for calculating an expression score for an annotated feature (cyan rectangle). Method I uses the sum of the number of reads crossing each nucleotide position (position scores) within the feature divided by the length of the feature. The position scores of the first 10 bp of the feature are shown above the sequence reads in the feature (horizontal black lines) at the position of every nucleotide (vertical red lines). Method II uses a similar approach but only uses an arbitrary portion of the 3′ end of the ORF, dividing the sum of reads crossing each position by the same arbitrary length. This method can be useful if the cDNA sequenced has only been oligo dT primed and so would contain a 3′–5′ bias. The third method simply calculates the number of reads that fall within the ORF, divided by the length of the ORF. The first five reads in the feature are shown numbered below the feature.

The quantification methods described above are limited by the need for accurate annotation of the reference genome that is used. This is particularly challenging for higher eukaryotes as the increasing transcriptional complexity of their genomes become more apparent. In the case of unannotated exons, the inclusion of such data could create variation between the expression level of genes surveyed using RNA-seq compared to microarrays. For instance, if the microarray probe for a given gene is in the 3' UTR, but the RNA-seq data for the same gene includes reads from a large, but rarely used, alternatively spliced exon, the length-normalized gene expression score for RNA-seq will be lower than the microarray value. Despite this potential complication, data from RNA-seq experiments have an important advantage over standard microarray data, in that it can be used to validate whether the novel transcribed regions are in fact novel exons of an annotated gene. By identifying short sequence reads which span exon–exon junctions ("trans-reads"), transcript isoforms can be revealed through the connections between exons. Initial efforts in this direction are very promising [10,11], however as short sequence reads cannot yet validate all the junctions in a given mRNA, the relative amounts of different isoforms must be derived from inferential methods.

Once expression scores have been calculated, it is often useful to visualize the data along with genome annotation information for specific regions. Based on RNA-seq results from model organisms such as fission yeast [16] and budding yeast [24], it is generally necessary to compress the values of nucleotide scores or reads counts by taking the log of the values (or similar transformation) in order to display the data clearly. Such illustration of the data, as shown in Fig. 3, has the added benefit of allowing fairly intuitive interpretation of the results in terms of identifying potential novel introns, novel genes or putative alternative poly adenylation sites.

7. Data storage, submission, archiving

The primary information generated from most next-generation sequencing approaches are either images or image related data, and as such, generally result in very large data files (see Table 2). While these images are stored locally for processing by the sequencing software, the enormous size of these files makes them impractical as a general method of storing the data. Indeed, as reagent costs decrease for the next-generation sequencing technologies, the cost of computer storage space for the data from a single run begins to approach the cost of re-sequencing the actual sample. While this regeneration of the data is not an ideal solution, it does highlight the problem caused by the sheer volume of data; a problem which is only likely to increase as new sequencing technologies advance.

Traditional DNA sequence repositories such as NCBI and EMBL are already being adjusted to allow for the storage of processed sequence data from next generation sequencing machines in the MINSEQE schema (Minimum Information about a high-throughput Nucleotide SEQuencing Experiment), analogous to the MIAME guidelines for microarrays [25]. As such, it will be possible to have access not only to final reads used for analysis, but also to the various data prior to filtering or normalization, in addition to complete descriptions of the method in which the RNA samples were collected and prepared before sequencing. Currently, the short read archive (NCBI) holds >5 trillion bases of next-generation sequencing data (deposited between Jun 2008 and Feb 2009), the bulk of which (96%) represents human data, mostly generated as part of the of the 1000 genomes project. Likewise, the European Read Archive (ERA), is mirroring the data in NCBI, while allowing direct submissions, in the same way that Genbank and EMBL have operated in the past.

8. Concluding remarks

The development of robust DNA sequencing technologies by Fredrick Sanger during the 1970s ushered in a new era in molecular biology, creating the ability to view the exact base pair composition of a gene [26,27]. Since this time, there have been numerous incremental technical improvements that have increased the throughput of sequencing machines using this methodology. Despite these advances, the easiest method to rapidly scale-up output was, until recently, the purchase of a larger number of sequencing machines. This lead to the creation of a relatively small number of large genome sequencing centers, operating like factories, and generating DNA sequence data from a selected list of organisms for the broader scientific community.

The arrival of next generation sequencing technologies has created the first break with this old sequencing paradigm, and can conceivably allow each university or research institute to generate DNA sequence on a scale that, in the past, would have only been possible at a genome sequencing center. The effects of this shift are wide ranging and multifaceted. In the first instance, the ability of *anyone* to generate massive amounts of cheap DNA sequence data from any organism has created a profound democratization of the process of genome sequencing. In the second instance, tech-

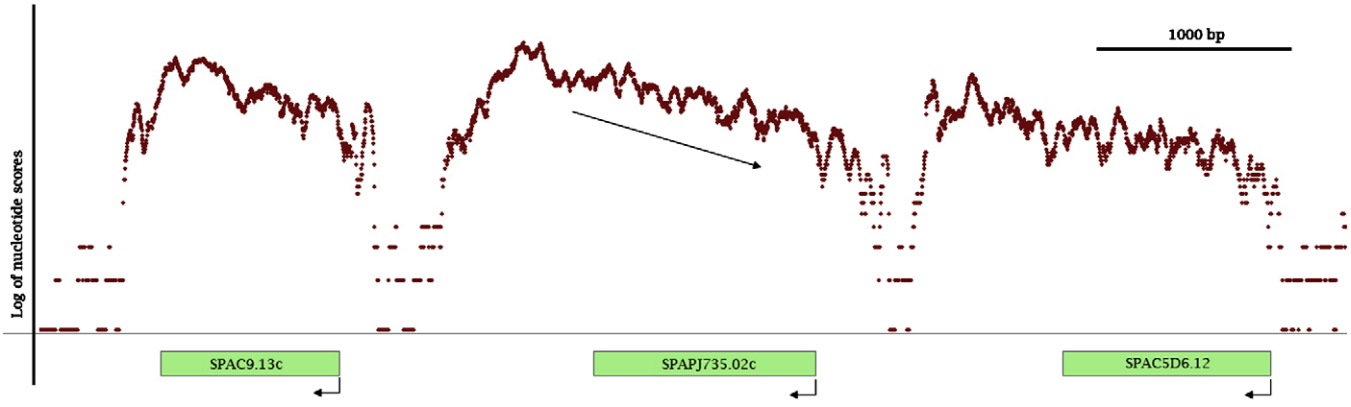

Fig. 3. An example of RNA-seq data plotted with corresponding annotation. The fission yeast genes SPAC9.13c, SPAPJ735.02c and SPAC5D6.12 with corresponding RNA-seq data [16] plotted along the *x*-axis while the *y*-axis represents the log (base 2) of the nucleotide scores (number of reads crossing every bp position in the region). The transcriptional orientation of the genes is shown by the small bent arrows, while the large slanted arrow indicates the 3′–5′ slope created in read frequency as a result of the reverse transcriptase falling off during RT. Note that some signals are present both inside and out of the ORF regions. Low level signal, while potentially a real biological phenomenon, could also be caused by contaminating gDNA in the RNA preparations. Such background can be removed from the displays/analysis by arbitrarily raising the floor (a value representing no transcription) of the data set, however this should be done only when justified through extensive additional controls.

nical applications, such as RNA-seq (and ChIP-seq, Meth-Seq, etc.) have evolved such that they now have the potential to dramatically improve, and therefore supplant, the previously used technology, microarrays, if the economics continue to improve.

While allowing powerful new applications like RNA-seq, the continually accelerating pace of technological change in the field of next generation sequencing also carries with it the real hazard of creating a glut of unused (or "under used") information. Substantial effort is already being invested in developing new approaches to deal with the volumes of data created by the current generation of new sequencing technologies, in order to maximize their potential benefit. The predicted output of other novel sequencing approaches, which could be available in the next few years, will dwarf even the output of current next-generation approaches. This raises the possibility that unless a great deal of effort is put into developing the computational tools and expertise to efficiently analyze the coming onslaught of data, there may eventually be thousands of sequenced genomes (and transcriptomes) just sitting on computer hard drives waiting to be analyzed.

Acknowledgments

The authors thank James Féthière and Martin Sauvageau for helpful discussion during preparation of the manuscript.

References

[1] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B.M. Ljung, J.W. Gray, D.G. Albertson, Nat. Genet. 20 (1998) 207–211.
[2] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, R.A. Young, Science 290 (2000) 2306–2309.
[3] V.R. Iyer, C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, P.O. Brown, Nature 409 (2001) 533–538.
[4] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, W. de Laat, Nat. Genet. 38 (2006) 1348–1354.
[5] R.E. Green, J. Krause, S.E. Ptak, A.W. Briggs, M.T. Ronan, J.F. Simons, L. Du, M. Egholm, J.M. Rothberg, M. Paunovic, S. Paabo, Nature 444 (2006) 330–336.
[6] A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, Cell 129 (2007) 823–837.
[7] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Science 316 (2007) 1497–1502.
[8] P.J. Campbell, P.J. Stephens, E.D. Pleasance, S. O'Meara, H. Li, T. Santarius, L.A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J.W. Teague, A. Menzies, I. Goodhead, D.J. Turner, C.M. Clee, M.A. Quail, A. Cox, C. Brown, R. Durbin, M.E. Hurles, P.A. Edwards, G.R. Bignell, M.R. Stratton, P.A. Futreal, Nat. Genet. 40 (2008) 722–729.
[9] J. Dostie, T.A. Richmond, R.A. Arnaout, R.R. Selzer, W.L. Lee, T.A. Honan, E.D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R.D. Green, J. Dekker, Genome Res. 16 (2006) 1299–1309.
[10] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, B.J. Blencowe, Nat. Genet. 40 (2008) 1413–1415.
[11] M. Sultan, M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, M.L. Yaspo, Science 321 (2008) 956–960.
[12] V. Nygaard, E. Hovig, Nucleic Acids Res. 34 (2006) 996–1014.
[13] R.N. Van Gelder, M.E. von Zastrow, A. Yool, W.C. Dement, J.D. Barchas, J.H. Eberwine, Proc. Natl. Acad. Sci. USA 87 (1990) 1663–1667.
[14] N. Cloonan, A.R. Forrest, G. Kolle, B.B. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel, A.J. Robertson, A.C. Perkins, S.J. Bruce, C.C. Lee, S.S. Ranade, H.E. Peckham, J.M. Manning, K.J. McKernan, S.M. Grimmond, Nat. Methods 5 (2008) 613–619.
[15] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Nat. Methods 5 (2008) 621–628.
[16] B.T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C.J. Penkett, J. Rogers, J. Bahler, Nature 453 (2008) 1239–1243.
[17] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C.C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K.C. Pang, J. Hallinan, J. Mattick, D.A. Hume, L. Lipovich, S. Batalov, P.G. Engstrom, Y. Mizuno, M.A. Faghihi, A. Sandelin, A.M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, C. Wahlestedt, Science 309 (2005) 1564–1566.
[18] R. Lister, R.C. O'Malley, J. Tonti-Filippini, B.D. Gregory, C.C. Berry, A.H. Millar, J.R. Ecker, Cell 133 (2008) 523–536.
[19] R. Li, Y. Li, K. Kristiansen, J. Wang, Bioinformatics 24 (2008) 713–714.
[20] H. Li, J. Ruan, R. Durbin, Genome Res. 18 (2008) 1851–1858.
[21] Z. Ning, A.J. Cox, J.C. Mullikin, Genome Res. 11 (2001) 1725–1729.
[22] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y.J. Chen, Nature 409 (2001) 860–921.
[23] R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S.E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M.R. Brent, D.G. Brown, S.D. Brown, C. Bult, J. Burton, J. Butler, R.D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A.T. Chinwalla, D.M. Church, M. Clamp, C. Clee, F.S. Collins, L.L. Cook, R.R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K.D. Delehaunty, J. Deri, E.T. Dermitzakis, C. Dewey, N.J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D.M. Dunn, S.R. Eddy, L. Elnitski, R.D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G.A. Fewell, P. Flicek, K. Foley, W.N. Frankel, L.A. Fulton, R.S. Fulton, T.S. Furey, D. Gage, R.A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T.A. Graves, E.D. Green, S. Gregory, R. Guigo, M. Guyer, R.C. Hardison, D. Haussler, Y. Hayashizaki, L.W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D.B. Jaffe, L.S. Johnson, M. Jones, T.A. Jones, A. Joy, M. Kamal, E.K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W.J. Kent, A. Kirby, D.L. Kolbe, I. Korf, R.S. Kucherlapati, E.J. Kulbokas, D. Kulp, T. Landers, J.P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D.R. Maglott, E.R. Mardis, L. Matthews, E. Mauceli, J.H. Mayer, M. McCarthy, W.R. McCombie, S. McLaren, K. McLay, J.D. McPherson, J. Meldrim, B. Meredith, J.P. Mesirov, W. Miller, T.L. Miner, E. Mongin, K.T. Montgomery, M. Morgan, R. Mott, J.C. Mullikin, D.M. Muzny, W.E. Nash, J.O. Nelson, M.N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M.J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K.H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C.S. Pohl, A. Poliakov, T.C. Ponce, C.P. Ponting, S. Potter, M. Quail, A. Reymond, B.A. Roe, K.M. Roskin, E.M. Rubin, A.G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M.S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J.B. Singer, G. Slater, A. Smit, D.R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J.P. Vinson, A.C. Von Niederhausern, C.M. Wade, M. Wall, R.J. Weber, R.B. Weiss, M.C. Wendl, A.P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R.K. Wilson, E. Winter, K.C. Worley, D. Wyman, S. Yang, S.P. Yang, E.M. Zdobnov, M.C. Zody, E.S. Lander, Nature 420 (2002) 520–562.
[24] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, M. Snyder, Science 320 (2008) 1344–1349.
[25] MGED (2008).
[26] F. Sanger, A.R. Coulson, J. Mol. Biol. 94 (1975) 441–448.
[27] F. Sanger, S. Nicklen, A.R. Coulson, Proc. Natl. Acad. Sci. USA 74 (1977) 5463–5467.
[28] M.N. Bainbridge, R.L. Warren, M. Hirst, T. Romanuik, T. Zeng, A. Go, A. Delaney, M. Griffith, M. Hickenbotham, V. Magrini, E.R. Mardis, M.D. Sadar, A.S. Siddiqui, M.A. Marra, S.J. Jones, BMC Genomics 7 (2006) 246.
[29] A.P. Weber, K.L. Weber, K. Carr, C. Wilkerson, J.B. Ohlrogge, Plant Physiol. 144 (2007) 32–42.
[30] D.J. Sugarbaker, W.G. Richards, G.J. Gordon, L. Dong, A. De Rienzo, G. Maulik, J.N. Glickman, L.R. Chirieac, M.L. Hartman, B.E. Taillon, L. Du, P. Bouffard, S.F.

Kingsmore, N.A. Miller, A.D. Farmer, R.V. Jensen, S.R. Gullans, R. Bueno, Proc. Natl. Acad. Sci. USA 105 (2008) 3521–3526.
[31] Y. Wei, S. Chen, P. Yang, Z. Ma, L. Kang, Genome Biol. 10 (2009) R6.
[32] L.J. Core, J.J. Waterfall, J.T. Lis, Science 322 (2008) 1845–1848.
[33] R. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones, M. Marra, Biotechniques 45 (2008) 81–94.
[34] R. Rosenkranz, T. Borodina, H. Lehrach, H. Himmelbauer, Genomics 92 (2008) 187–194.
[35] F. Denoeud, J.M. Aury, C. DaSilva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli, O. Jaillon, F. Artiguenave, Genome Biol. 9 (2008) R175.
[36] H. Li, M.T. Lovci, Y.S. Kwon, M.G. Rosenfeld, X.D. Fu, G.W. Yeo, Proc. Natl. Acad. Sci. USA 105 (2008) 20179–20184.
[37] R.D. Morin, M.D. O'Connor, M. Griffith, F. Kuchenbauer, A. Delaney, A.L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst, C.J. Eaves, M.A. Marra, Genome Res. 18 (2008) 610–621.
[38] A. Grimson, M. Srivastava, B. Fahey, B.J. Woodcroft, H.R. Chiang, N. King, B.M. Degnan, D.S. Rokhsar, D.P. Bartel, Nature 455 (2008) 1193–1197.
[39] C.A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, A.M. Chinnaiyan, Nature 458 (2009) 97–101.
[40] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, Y. Gilad, Genome Res. 18 (2008) 1509–1517.