BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 2. 181–197
Advance Access published on 27 October 2009
doi:10.1093/bib/bbp046

Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing

David Stephen Horner, Giulio Pavesi, Tiziana Castrignanò, Paolo D'Onorio De Meo, Sabino Liuni, Michael Sammeth, Ernesto Picardi and Graziano Pesole

Submitted: 3rd August 2009; Received (in revised form): 6th September 2009

Abstract

Technical advances such as the development of molecular cloning, Sanger sequencing, PCR and oligonucleotide microarrays are key to our current capacity to sequence, annotate and study complete organismal genomes. Recent years have seen the development of a variety of so-called 'next-generation' sequencing platforms, with several others anticipated to become available shortly. The previously unimaginable scale and economy of these methods, coupled with their enthusiastic uptake by the scientific community and the potential for further improvements in accuracy and read length, suggest that these technologies are destined to make a huge and ongoing impact upon genomic and post-genomic biology. However, like the analysis of microarray data and the assembly and annotation of complete genome sequences from conventional sequencing data, the management and analysis of next-generation sequencing data requires (and indeed has already driven) the development of informatics tools able to assemble, map, and interpret huge quantities of relatively or extremely short nucleotide sequence data. Here we provide a broad overview of bioinformatics approaches that have been introduced for several genomics and functional genomics applications of next-generation sequencing.

Keywords: *short-read alignment; RNA-Seq; ChIP-Seq; single nucleotide polymorphism; editing; epigenomics*

INTRODUCTION

The introduction of next-generation sequencing (NGS) technologies has had a huge impact upon genomics and functional genomics, indeed these methods are rapidly supplanting the conventional Sanger, or di-deoxy terminator, strategy [1] that

Corresponding author. Graziano Pesole. Tel: +39-080-5443588; Fax: +39-080-5443317; E-mail: graziano.pesole@biologia.uniba.it
David Stephen Horner is an assistant professor of Molecular Biology at the University of Milan (Italy). His research interests include comparative genomics, non-coding RNA and molecular phylogenetics.
Giulio Pavesi is an assistant professor of Computer Science at the University of Milan (Italy). His research interests are mainly focused on bioinformatics in general, and regulatory motif discovery in particular. He also works on discrete models of complex systems.
Tiziana Castrignanò is the leader of bioinformatics at CASPUR (Consorsio per le Aplicazioni di Supercalcolo per l'Universita' e Ricerca). She received her PhD in Biophysics in 1999 from the University of Rome 'La Sapienza'. Her primary research interests are the development of high performance bioinformatics services and databases. She has provided technological support in many national and international bioinformatics research projects.
Paolo D'Onorio De Meo got a Bachelors Degree in Computer Science at the University of Rome 'La Sapienza' in 2004. Since 2004 he is bioinformatics developer at CASPUR (Consorsio per le Aplicazioni di Supercalcolo per l'Universita' e Ricerca).
Sabino Liuni is a senior technologist at the Institute of Biomedical Technology (National Research Council) in Bari (Italy). His research interests in the field of bioinformatics regard computational methods for the analysis of next-generation sequencing data,
Michael Sammeth is a researcher in the Research Unit on Biomedical Informatics at the Centre of Genomics Regulation (CRG) and University Pompeu Fabra, Barcelona (Spain). His research interests are in the field of Bioinformatics, and particularly in the analysis of alternative splicing.
Ernesto Picardi is an assistant professor of Molecular Biology at the University of Bari (Italy). His research interests are mainly focused on bioinformatics and molecular evolution.
Graziano Pesole is a full professor of Molecular Biology at the University of Bari (Italy) leading a research team in 'Bioinformatics and Comparative Genomics' at the Institute of Biomedical Technology (National Research Council). His research interests include bioinformatics, development of tools for genome annotation, comparative genomics and molecular evolution.

has, in various manifestations, been the principal method of sequencing DNA since its inception in the late 1970s.

The development of these new massively parallel sequencing technologies has sprung from recent advances in the field of nanotechnology, from the availability of optical instruments capable of reliably detecting and differentiating millions of sources of light or fluorescence on the surface of a small glass slide and from the ingenious application of classic molecular biology principles to the sequencing problem. Another important consideration is that, in the context of an already available genome sequence, many problems—such as the identification of single nucleotide polymorphisms (SNPs)—need not require the generation of ever longer sequence reads, because most possible 'words' of length >25 or 30 only occur at most once even in relatively large genomes—allowing, for the most part, unambiguous assignment of even the shortest reads to a locus of origin in a reference genome. Thus, available NGS technologies produce large numbers of short sequence reads and are typically used in 'resequencing' applications, implying the availability of a reference sequence identical, or highly similar, to the source of the genetic material under consideration.

In addition to the conventional objectives of genome resequencing/SNP discovery, the characteristics of these technologies permit them to be efficiently applied to a number of other applications. For example, NGS of cDNA can be used to provide a comprehensive snapshot of the transcriptome, facilitating gene annotation and identification of splicing variants. These novel technologies have also been extensively applied to the characterization of small RNA populations, the identification of microRNA targets in plants, the characterization of genomic regions bound by transcription factors (TFs) and other DNA binding proteins, the identification of genome methylation patterns, the characterization of RNA editing patterns and metagenomics projects [2–9]. It is likely that a series of other applications for NGS methods will be unveiled within the next years. Currently available next-generation sequencers rely on a variety of different chemistries to generate data and produce reads of differing lengths, but all are massively parallel in nature and present new challenges in terms of bioinformatics support required to maximize their experimental potential.

In this review, we will not attempt to provide a detailed description of the sequencing technologies themselves, interested readers are referred in particular to several excellent recent reviews [4,10,11]. Rather, we will touch upon some of the applications for these technologies that have emerged in genomics and functional genomics research [6, 12], focusing particularly on bioinformatics tools that have been developed for data management and analysis. Given the rate of development in this field, we will not attempt to mention every instrument that has been presented, rather, we will try to provide a general overview of trends and focus on tools with which the authors of this review have first-hand experience.

NGS PLATFORMS

Three distinct NGS platforms have already attained wide diffusion and availability. Some characteristics of their throughput, read-lengths and costs (at the time of writing) are presented in Table 1. A common thread for each of these technologies over the last years has been continuous improvement in performance (increased numbers and lengths of reads and consequent reduction in costs per base sequenced), it is therefore anticipated that the figures provided will rapidly become outdated, however, they serve to illustrate that the Roche 454 technology [13] already provides a realistic substitute for many applications of conventional Sanger sequencing at greatly reduced cost, while the Illumina Genome Analyser [14] and ABI SOLiD [15] platforms generate an order of magnitude more reads of (relatively) reduced length, characteristics that, as we will see, render them, for now, more suitable for other applications.

The aforementioned methods all rely on a template amplification phase prior to sequencing. However, the available Helicos technology [16] avoids the amplification step and provides sequence data for individual template molecules, minimizing the risk of introducing substitutions during amplification. In principle bioinformatics approaches developed for the analysis of data generated by the Illumina GA and ABI SOLiD platforms should also be suitable for data generated by the Helicos method, as all three platforms provide reads of comparable lengths. Finally, other methods, based on either nanopore technology or tunneling electron microscopy have been proposed (for reviews see [17–19]).

Table I: Performances and features of the major next-generation sequencing platforms (single-end reads)

Technology	Roche 454			Illumina		ABI Solid		
Platform	*GS 20*	*FLX*	*Ti*	*GA*	*GA II*	*1*	*2*	*3*
Reads (*M*)	0.5	0.5	1	28	100	40	115	400
Read length	100	200	350	35	75	25	35	50
Run time (*d*)	0.2	5	0.3	0.4	4.5	6	5	6–7
Images (TB)	0.01	0.01	0.03	0.5	1.7	1.8	2.5	3

Detailed information on the performance of such approaches is not yet available, although it is hoped that they could yield individual reads of lengths measured in megabases. Given that such methods remain broadly inaccessable at the present time, and that the nature of data generated should be fundamentally different from those provided by available platforms, potential bioinformatics developments connected to these methods are considered to be beyond the scope of the current review.

While both the Illumina Genome Analyser and Roche 454 platforms use innovative techniques to amplify and sequence template molecules, they share the underlying principle of 'sequencing by extension' used in the Sanger methodology. This is to say that single bases, complementary to the template molecule are sequentially added to a nascent strand and their identity determined by chemical means. However, the ABI-SOLiD sequencing technology uses a unique chemistry whereby oligonucleotides complementary to a series of bases in the sequencing template are ligated to a nascent molecule and the identity of the fist two bases of the ligated oligonucleotide is specified by a degenerate four color code (each color specifies four different dinucleotides). This approach provides some benefits in terms of accuracy since each base in the template is interrogated twice in independent primer rounds. As a consequence, color reads can be translated into base reads only if the first base of the sequence—or more commonly the last base of the primer used—is known (although see the section on mapping tools). In resequencing applications, careful consideration of SOLiD sequencing data can allow differentiation between sequencing errors and biological SNPs. Effectively, sequence errors cause major changes in all downstream bases (while variation between template and reference sequence cause single mismatches in mapped reads).

Analogously to automated Sanger sequencing, NGS platforms provide quality values or quality scores describing the likelihood that a base call is incorrect. The *phred* algorithm [20] assigns a quality value for each base in a Sanger read in which larger numbers designate smaller error probabilities. A Q20 value, for example, corresponds to a 1 in 100 error probability, and a Q30 value to a 1 in 1000 error rate. NGS platforms have different error profiles and, thus, quality values need to be derived accordingly. In Illumina GA, the meaning of the quality values is relatively close to capillary sequencers. Moreover, Illumina scores are asympotically identical at higher quality values. Sanger *phred* quality scores range from 0 to 93 (using ASCII 33–126 in fastq), whereas Illumina quality values range from −5 to 40 (using ASCII 59–104 in fastq) or from 0 to 40 (using ASCII 64–104 in fastq) depending on Illumina GA version (1.0 in the first case and 1.3 in the second case). In the SOLiD system, quality scores are assigned to each color and calculated using a *phred* like score $q = -10 \times \log 10(p)$, where p is the predicted probability that the color call is incorrect. SOLiD quality values generally range from 0 to 45, although the exact relationship between color scores and *phred* values is not completely known. For 454 reads, quality values per base range from 0 to 40 and also in this case they are calculated using a *phred* like algorithm. However, the error probability in 454 reads is mainly related to the probability that a base is an overcall. Roche 454 reads are prone to insertion and deletion errors rather than miscalling errors more frequent in Sanger reads. Furthermore, a variety of relatively simple, text-based file formats are used by different NGS platforms. For a discussion of these formats and up-to-date discussions on the implications of quality scores, readers are referred to an excellent web based discussion forum (http://seqanswers.com).

Mapping strategies

The first and arguably most crucial step of most NGS analysis pipelines is to map reads to sequences

of origin. The occurrence of nucleotide polymorphisms between reference genome and sampled individuals, relatively high rates of sequencing errors, RNA editing and epigenetic modifications all require efficient mapping with limited numbers of mismatches (typically two or three in a 35 base alignment) and potentially single-base insertions or deletions. Specific strategies are needed for mapping spliced transcript sequences to genome sequences. Statistically speaking, reads of 30 bp should be expected to yield unique matches on most genomes, In practice, some reads do not map anywhere on the genome—owing to DNA contamination or sequencing artifacts, while some map exactly or approximately at multiple positions, as a result of the complex and repetitive nature of genome sequences—potentially reducing the effective output of NGS platforms.

Mapping of reads is a distinctive manifestation of perhaps the oldest bioinformatics problem, sequence alignment. However, classical methods such as pure Smith–Waterman dynamic programming, or indexing of longer *k*-mers in the template sequence (BLAT) [21], or combinations of the two (e.g. BLAST) [22] are not well suited to the alignment of very large numbers of short sequences to a reference sequence [23].

To avoid the need for expensive dedicated hardware, the overall goal of short read mapping is to obtain satisfactory results as efficiently (in terms of time and memory requirements) as possible. As a result, many methods are based on the similar principles and algorithms, but differ in the 'programming tricks' or ad hoc heuristics used to increase speed at the price of minimal loss of accuracy. Research in this field is booming and new, or modified mapping tools currently appear on an almost weekly basis [24]. Thus, here we will confine ourselves to a description of the general principles underlying the most successful algorithms, and very brief descriptions of a few—leaving to the interested reader the task of keeping abreast with one of the hottest and most rapidly growing fields of modern bioinformatics.

The principle of creating an index of the positions of all distinct *k*-mers in either the sequence reads or the genome sequence underlies most short read mapping tools. Applied to our problem, suppose that we have to map a tag of length 32 with up to two mismatches. The tag can be defined as the concatenation of four substrings of length 8 bp. Since at most two mismatches are allowed, then at least two of the substrings are guaranteed to match exactly the genome. The matching substrings can be adjacent, or separated by one or two mismatching substrings. Thus, if we want to build an index for the genome, we can index substrings of length 16 in three separate ways, corresponding to (i) two adjacent substrings of 8 bp, or (ii) the concatenation of two substrings of 8 bp separated by 8 bp, or (iii) the concatenation of two substrings of 8 bp spaced by 16 bp. These combinations of substrings are used as seeds for the initial exact matching stage, since in case of a tag matching the genome with up to two substitutions we will find an exact match for two 8 bp substrings of the tag in one of the three indices. Alternatively, in the same situation we can be certain that at least a substring of length 11 will match exactly. Hence, we can index substrings of length 11 and employ them as initial matching seed. In both cases, once a seed has been matched, it can be extended, allowing for mismatches and/or insertion and deletions. The choice of which dataset (reads or reference) is indexed can have significant implications upon speed and memory requirements. Essentially, indexing the larger dataset will require more memory, but will accelerate the mapping phase. While the size of a large eukaryotic genome (such as those of mammals or many higher plants) is measured in billions of base pairs, the overall size of the set of tags to be mapped can now be of a similar magnitude (20 billion bases per run for the ABI SOLiD 3 system). Thus, building an index for the genome and matching the tags against the index can often present similar memory requirements to the inverse operation (indexing the tags and matching the genome against them). However, the latter approach has the benefit of being trivially scalable: if the available memory is not sufficient to hold the index of the whole set of tags to be processed, then the tags can be split into subsets and each subset can be processed separately (or in parallel, if several computing cores are available). The final result is obtained by merging of the results obtained for each subset. While the computation time is increased, tools of this kind can be used with standard personal computers.

The most fundamental differences between available mapping algorithms are, arguably, whether the genome or the sequence reads are indexed, and the indexing method applied. Additionally, different methods may or may not allow the presence of indels

in alignments, the reporting of only unique best matches or of all matches within a defined maximum Hamming—or edit—distance. As mentioned previously, various heuristics have also been introduced to accelerate searches, for example 'quality scores' indicating the confidence of base calls can be used to limit the search space. Thus, mismatches can be confined only to those tag nucleotides that are deemed to be 'less reliable', or reads containing low-quality base calls can simply be excluded. Alternatively, since less reliable base calls are often located near the end of reads, one could require exact matching for the beginning of the reads and allow for mismatches in the rest. When entire tags do not generate a satisfactory mapping, the last bases (more likely to include sequence errors) can be trimmed away and the matching can be repeated for the shorter reads.

Many mapping tools have been reported, some of them have been designed for a specific sequencing platform and others are more general purpose (Table 2). A commercial aligner called ELAND was developed in parallel with the Solexa sequencing technology, and it is provided for free for research groups that buy the sequencer. Probably the first tool introduced for this task, ELAND indexes the tags, and is based on the aforementioned tag-splitting strategy allowing mismatches. It is still one of the fastest and less memory-greedy pieces of software available. Likewise, SeqMap [25] builds an index for the reads by using the longest substring guaranteed to match exactly, and scans the genome against it. It allows the possibility of insertions and deletions in alignments. ZOOM [26] is also based on the same principles as ELAND, with the difference that reads are indexed by using 'spaced' seeds that can be denoted with a binary string. For example, in the spaced seed 111010010100110111, 1's mean a match is required at that position, 0's indicate 'don't care' positions. Only positions with a '1' in the seed are indexed. The performance reported is faster than ELAND, at the price of higher memory requirements.

Short Oligonucleotide Alignment Program (SOAP) [27] was one of the first methods published for the mapping of short tags, in which both tags and genome are first of all converted to numbers using 2-bits-per-base encoding. To admit two mismatches, a read is split into fragments as in ELAND. Mapping with either mismatches or indels is allowed. Since for technical reasons reads always exhibit a much higher number of sequencing errors at the 3′-end, which sometimes make them unalignable to the genome, SOAP can iteratively trim several bases from the 3′-end and redo the alignment until hits are detected or the remaining sequence is too short for specific alignment. The main drawback is the memory requirement, reported to be >10 GB for the human genome.

PASS [28] holds the hash table of the genomic positions of seed substrings (typically 11 and 12 bases) in RAM memory as well as an index of precomputed scores of short words (typically seven and eight bases) aligned against each other. The program matches each tag performing three steps: (i) it finds matching seed words in the genome; (ii) for every match checks the precomputed alignment of the short flanking regions (thus including insertions and deletions); and (iii) if step 2 is passed, it performs an exact dynamic alignment of a narrow region around the initial match. The performance is reported to be much faster than SOAP, but once again at the price of high memory requirements (10's of GB) for the genomic index.

The maximum oligonucleotide mapping (MOM) [29] algorithm searches for exactly matching short subsequences (seeds) between the genome and tag sequences, and performs ungapped extension on those seeds to find the longest possible matching sequence with a user specified number of mismatches. To search for matching seeds MOM creates a hash table of subsequences of fixed length k (k-mers) from either the genome or the tag sequences, and then sequentially reads the un-indexed sequences searching for matching k-mers in the hash table. As in SOAP, tags which cannot be matched entirely given a maximum number of errors are automatically trimmed. The performance reported is better than SOAP, in terms of number of tags successfully matched. Again, more than 10 GB of memory are needed for typical applications.

The space requirements of building a genomic index with a hash table can be reduced by using more efficient strategies. A good (at least theoretically) performance is also obtained by *vmatch* [30], which employs enhanced suffix arrays for a number of different genome-wide sequence analysis applications.

Bowtie [31] (and a newer version of SOAP [32]) employ a Burrows–Wheeler index based on the full-text minute-space (FM) index, which has a reported

Table 2: Tools for the analysis of next-generation sequencing data in several application categories

Tool	Website	Category	Platform[a]
ELAND	http://www.illumina.com/pages.ilmn?ID=315	Alignment	GA
Soap	http://soap.genomics.org.cn	Alignment	GA
ZOOM	http://www.bioinfor.com	Alignment	GA, SO
PASS	http://pass.cribi.unipd.it	Alignment	GA, SO, GS
MOM	http://mom.csbc.vcu.edu	Alignment	GA
Vmatch	http://www.vmatch.de/	Alignment	GA
Bowtie	http://bowtie.cbcb.umd.edu	Alignment	GA
CloudBurst	http://cloudburst-bio.sourceforge.net/	Alignment	GA
BWA	http://maq.sourceforge.net/bwa-man.shtml	Alignment	GA
SHRiMP	http://compbio.cs.toronto.edu/shrimp/	Alignment	GA, SO
AB mapreads	http://solidsoftwaretools.com/gf/project/mapreads/	Alignment	SO
MuMRescueLite	http://genome.gsc.riken.jp/osc/english/dataresource/	Alignment	SO
MAQ	http://maq.sourceforge.net	Alignment	GA, SO
SeqMap	http://biogibbs.stanford.edu/~jiangh/SeqMap/	Alignment	GA
RMAP	http://rulai.cshl.edu/rmap/	Assembly	GA
FindPeaks	http://www.bcgsc.ca/platform/bioinfo/software/findpeaks	ChipSeq analysis	GA, SO, GS
F-Seq	http://www.genome.duke.edu/labs/furey/software/fseq	ChipSeq analysis	GA
SISSRS,	http://sissrs.rajajothi.com/	ChipSeq analysis	GA
QuEST	http://www.stanford.edu/~valouev/QuEST/QuEST.html	ChipSeq analysis	GA
MACS	http://liulab.dfci.harvard.edu/MACS/	ChipSeq analysis	GA
Chipseq.peak finder	http://woldlab.caltech.edu/html/software	ChipSeq analysis	GA
ChIPDiff	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperChIPDiff.htm	ChipSeq analysis	GA
CisGenome	http://www.biostat.jhsph.edu/~hji/cisgenome/	ChipSeq analysis	GA
G-Mo.R-Se	http://www.genoscope.cns.fr/externe/gmorse/#Download	Gene annotation	GA
UEA plant sRNA toolkit	http://srna-tools.cmp.uea.ac.uk/	General smallRNA tools	GA
mirDeep	http://www.mdc-berlin.de/rajewsky/miRDeep	mirRNA identification	GA, SO, GS
MirCat	http://srna-tools.cmp.uea.ac.uk/	mirRNA identification	GA
QSRA	http://qsra.cgrb.oregonstate.edu/	short read assembly	GA
ALLPATHS	http://www.broadinstitute.org/science/programs/genome-biology/computational-rd/computational-research-and-development	short read assembly	GA
Velvet	http://www.ebi.ac.uk/~zerbino/velvet/	short read assembly	GA
EDENA	http://www.genomic.ch/edena.php	short read assembly	GA
VCAKE	http://mac.softpedia.com/get/Math-Scientific/VCAKE.shtml	short read assembly	GA, SO, GS
SHARCGS	http://sharcgs.molgen.mpg.de/download.shtml	short read assembly	GA
EULER-SR	http://euler-assembler.ucsd.edu/portal/	short read assembly	GA
SSAKE	http://www.bcgsc.ca/platform/bioinfo/software/ssake/releases/3.2	short read assembly	GA, SO, GS
CLEAVELAND	http://www.bio.psu.edu/people/faculty/Axtell/AxtellLab/Software.html	smallRNA target identification (plants)	GA
POLYBAYES	http://bioinformatics.bc.edu/marthlab/PolyBayes	SNP calling	GS, SO, GS
SLIDER	http://www.bcgsc.ca/platform/bioinfo/software/slider	SNP calling	GA, SO, GS
QPalma	http://www.fml.tuebingen.mpg.de/raetsch/suppl/qpalma	Spliced Read Mapping	GA
Tophat	http://tophat.cbcb.umd.edu/	Spliced Read Mapping; transcript quantification	GA
Erange	http://woldlab.caltech.edu/rnaseq/	Spliced Read Mapping; transcript quantification	GA
FluxCapacitor	http://flux.sammeth.net/	Transcript quantification	GA, SO, GS

A web version of this table, which is continuously updated, can be found at http://mi.caspur.it/ngs/software_review.php.
[a]GA, Illumina; SO, AB SOLID; GS, Roche 454 FLX.

memory requirement of only ~1.3 GB for the human genome. In this way, Bowtie can run on a typical desktop computer with 2 GB of RAM. However, if one or more exact matches exist for a tag, Bowtie always reports them, but if the best match is an inexact one then Bowtie is not guaranteed in all cases to find it. BWA [33] is also based on the Burrows–Wheeler transform.

Mapping and Assembly with Quality (MAQ [34]), one of the most successful tools in this field, is specifically devised to take advantage of the nucleotide-by-nucleotide 'quality scores' that come together with the Illumina reads. The idea is that mismatches due to errors in sequencing should mostly appear at those positions in the tags that have a 'low-quality' score, while those due to SNPs should always appear at the same position in the genomic sequence. Mismatches are thus 'weighted' according to their respective quality scores. By default, six hash tables are used, ensuring that a sequence with two mismatches or fewer will be hit in an ELAND-like fashion. The six hash tables correspond to six spaced seeds analogous to that used in ZOOM. By default, MAQ indexes the first 28 bp of the reads. While very fast, MAQ is based on a number of heuristics that do not always guarantee to find the best match for a read.

Sequence quality scores provided by Illumina are also employed by RMAP [35] wherein positions in reads are designated as either high- or low-quality. Low-quality positions always induce a match (i.e. act as wild-cards). To prevent the possibility of trivial matches, a quality control step eliminates reads with too many low-quality positions (a similar filter has also been implemented in PASS see above). CloudBurst [36] is a RMAP-like algorithm that supports cloud computation using the open-source Hadoop implementation of MapReduce to parallelize execution using multiple nodes.

Some mapping tools including MAQ [34], BWA [33], PASS [28], SHRiMP [37] and AB Mapreads (Zhang *et al.*, unpublished) work within color space—both for the reference sequence and reads. In this way, it is possible to employ conventional alignment algorithms that have been developed for Illumina GA and Roche 454 short reads.

The performance of the different methods can be measured according to different parameters: time required, memory occupation, disk space and in case of heuristic tools, the actual number of reads that have been assigned correctly to their original position on the genome. In turn, the choice of a given method against another one depends on how many tags have to be mapped, and quite naturally to the specifics of the computing equipment available.

Mapping simulation

To illustrate variation in performance of different short-read mapping tools and to highlight some likely complications of the nature of NGS data—we have performed two simple studies to simulate the RNA-seq transcriptome sequencing approach. We selected three programs (SOAP, BOWTIE and PASS) that we have previously used in our group, including PASS because it is one of the few mapping tools supporting MS Windows. In all experiments, parameters were adjusted such that each program should find all equally best matches to the reference sequence, with up to 2 mismatches. First, we performed a naive experiment where we randomly generated ~4-million 35 base long reads from annotated human transcripts dataset (RPM). PolyA tails were not simulated [38]. These reads were mapped to both the human transcriptome from which the reads originated (and which should provide perfect matches to each read) and to the human genome sequence from which the transcriptome originated (to which reads not covering splice junctions should give perfect matches). Table 3 shows various statistics regarding the speed, memory use and sensitivity of each of these mapping tools in this first simulation. We see clearly that SOAP provides the fastest performance while Bowtie uses marginally less RAM, with Bowtie correctly mapping 99.99% of all reads and a marginally lower mapping rate for SOAP. The apparently poor performance of PASS is likely due to imperfect parameterization and failure to identify all map positions for reads matching on a large numbers of transcripts.

We next considered the accuracy of the same instruments when mapping reads to the complete human genome (anticipating that reads derived from splice junctions will for the most part not map to their correct loci of origin). In this case, all methods map around 75% of the reads correctly. The decay from the previous scenario is due principally to the fact that reads spanning splice junctions (11.7% of all reads) tend not to map correctly to the genome sequence.

To provide a more realistic simulation, we employed a modified version of the Flux Simulator software (http://flux.sammeth.net/) which allows

Table 3: Results of the mapping simulation using randomly generated perfect match (RPM) and randomly-generated error-containing (RER) reads against the original transcriptome or the complete genome (hg18)

Simulated data versus reference sequence		Program		
		Bowtie	**PASS**	**SOAP**
RPM versus Transcriptome	Reads mapped	3 995 190	3 995 185	3 975 019
(3 995 721 reads)		(99.99%)	(99.99%)	(99.48%)
	Reads mapped correctly	3 995 190	3 562 888	3 966 646
		(99.99%)	(89.17%)	(99.27%)
	RAM required	160 MB	2.08 GB	1.40 GB
	Total processor time	244.87 s	346.76 s	86.11 s
RPM versus Genome	Reads mapped	32 988 443	3 380 343	3 300 773
		(82.55%)	(84.60%)	(82.61%)
(4 000 000 reads, 469 577 spliced)	Reads mapped correctly	3 034 232	3 066 025	2 991 559
		(75.94%)	(76.73%)	(74.87%)
	RAM required	1.24 GB	12.96 GB	2.56 GB
	Total processor time	255.9 s	1928.0 s	78.6 s
RER versus Transcriptome	Reads mapped	4 168 549	4 183 679	4 058 196
		(90.52%)	(90.85%)	(88.13%)
(4 604 890 reads)	Reads mapped correctly	3 987 222	3 259 096	3 833 970
		(86.59%)	(70.77%)	(83.26%)
RER versus Genome	Reads mapped	3 607 856	3 812 898	3 608 220
		(78.35%)	(82.80)	(78.36%)
(4 604 890 reads, 530 092 spliced)	Reads mapped correctly	3 497 369	3 503 184	3 359 094
		(75.95%)	(76.08%)	(72.95%)
	Incorrectly mapped spliced reads	15 050	80 089	14 324
		(2.84%)	(15.11%)	(2.70%)

We used the programs Bowtie (v0.99.3), PASS (v 0.71) and SOAP (v 2.16) fixing parameters for allowing up to two mismatches. The data used for the simulation can be found at http://mi.caspur.it/shortreads/download/.

the simulation of NGS transcriptome data under detailed models of cDNA synthesis, nebulization, size fractionation, variation in transcription start and polyadenylation sites and the introduction of sequencing errors under user defined models, in this case an empirically deduced model for Illumina sequencing [39]. In total, 4 604 890 36-bp long reads were simulated, of which 530 092 covered splice junctions (dataset RER). Again all methods were used to map reads to both the transcriptome and genomic sequences of origin (Table 3). Here a more complicated picture is observed. Against the transcriptome, Bowtie and SOAP correctly mapped 86.6 and 83.3% of reads, respectively, while PASS showed inferior performance. In total, 11.5% of reads recovered consisted at least partially of polyA tails and were not mapped to the transcriptome dataset as polyA tails were excluded from the transcripts for computational reasons. Thus in our simulation, sequence errors, mis-priming and alternative polyadenylation seem to have little effect on our capacity to correctly map reads with Bowtie and SOAP. When the reads were mapped to the genome, correct placements fell to ~75% for all methods, the decay corresponding well to the known proportion of reads covering splices. Nevertheless, the differences observed between methods illustrate that, particularly when error prone short reads are mapped to genomic sequences, a substantial number of artifactual placements are generated (mostly due to the presence of sequencing errors) and that the different heuristics used by different algorithms can find different imperfectly matching map positions.

Metagenomics and the *de novo* assembly of short sequence reads

Until now, we have considered applications that rely on mapping next-generation sequence data to available reference genome sequences. However, at least for smaller bacterial genomes, even the shortest reads can be used to effectively assemble genome sequences *de novo*, and even where complete closure of the genome is not possible, large contigs can be reliably constructed from such data provided that repeated sequences are not overly abundant. It should be noted that the continued increase in length

of reads obtained by NGS platforms suggest that in the near future, *ab initio* sequencing of some eukaryotic genomes with technologies such as Illumina or ABI SOLiD is likely to become a realistic prospect, while near-complete drafts of many microbial genomes can now be produced using the 454 technology [13,40,41].

Furthermore, the fields of metagenomics and microbial community analysis are not immune to the lure of NGS, although unsurprisingly, until now, the 454 technology has been the most widely used for these applications [42]. As with aligners, progress in the development of *de novo* assemblers has been rapid and is ongoing, and as with the aforementioned tools, at least most of the available *de novo* short read assemblers utilize a common underlying technique. As with the mapping problem, different tools use different algorithms to choose 'optimal' solutions and thus generate putative contigs. In the context of metagenomics, after contig assembly, high-throughput identification and phylogenetics strategies are required for the reconstruction of microbial communities, for a recent review of this field, interested readers are directed towards a review by Petrosino *et al.* [42].

Tools for whole-genome shotgun fragment assembly of conventional sequence data such as Atlas [43], ARACHNE [44], PCAP [45] and Phusion [46] are not able to handle the larger numbers of reads produced by NGS platforms or the higher error frequencies in these reads. However, they have proved useful for the development of *de novo* genome assemblers.

Currently available applications for *de novo* assembly of NGS data include: QSRA [47], ALLPATHS [48], Velvet [49], EDENA [50], VCAKE [51], SHARCGS [52], EULER-SR [53], SSAKE [54]. VCAKE, SSAKE and Velvet use De Bruijn graphs [55] to summarize the distribution of overlapping reads, while EULER-SR employs an alternative approach.

Such approaches tend to require a trade-off between production of the fewer long contigs with lower overall genomic coverage, or a higher number of shorter contigs with a higher overall genomic coverage.

Comparisons of QSRA with EDENA, Velvet, SSAKE and VCAKE algorithms indicated that QSRA was much faster than the other tools [47]. In these tests, EDENA and VELVET yielded the longest contigs with lower genomic coverage, while QSRA, SSAKE and VCAKE generally produce a higher number of shorter contigs, while QSRA gave the highest genomic coverage.

The Short-read Assembler based on Robust Contig extension for Genome Sequencing (SHARCGS) algorithm is capable of assembling millions of very short reads and manages sequencing errors. The performance of this algorithm was evaluated against SSAKE and EULER-SR [52]. It seems that SSAKE is particularly vulnerable to the presence of sequencing errors, while all contigs generated by SHARCGS were identical to the source sequences. EULER-SR is time-consuming, particularly when many sequencing errors render the graphs complex and EULER-SR runs into performance problems. Finally, the ALLPATHS algorithm allows the analysis of paired reads and unpaired reads for *de novo* assembly of whole genome shotgun microreads (25–50 bases). For a detailed view of technical and algorithmic issues in *de novo* assembly of short reads, readers are referred to [56].

Detection of SNPs and editing sites by NGS technologies

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation in humans and a resource for mapping complex genetic traits [57] as they can alter DNA, RNA and protein sequences at different levels [58]. SNPs are often identified using data from high-throughput sequencing projects and reads are typically aligned to the corresponding genomic reference and sequencing errors are discerned from genetic variations using quality scores as additional guidance [59]. The probability that an inferred SNP should be real can be assessed using Bayesian inference statistics implemented in tools such as POLYBAYES [59]. NGS platforms can improve the detection accuracy of SNPs thanks to increased sequencing depth. To date, all NGS technologies have been used to infer SNPs in mammalian genomes although platforms ensuring deep coverage (Illumina and SOLiD), have been preferred. Recently, a study by Smith and colleagues [60] showed that single mutations could be reliably detected given at least 10–15-fold nominal sequence coverage. Among high-throughput strategies SOLiD seems preferable for this purpose since its color-space system can discern sequencing errors from genuine variations [37, 61], while the Slider software uses an innovative

approach whereby Illumina reads are considered only as a series of quality scores and used to generate the most probable sequence to be mapped onto a reference [62]. After alignment, base positions with a high probability to be SNPs are selected. Slider also introduces an elegant Illumina base caller. In contrast, Roche 454 reads provide lower coverage per base but with high quality. However, pyrosequencing can introduce biases in SNP detection when homopolymeric strings are present. Tools like Pyrobayes can overcome such limitations in 454 data by improving the base calling and employing the Bayesian inference to identify SNPs after an *ad hoc* read mapping that uses modified gap costs to handle potential errors at homopolymeric stretches [63].

The increased availability of paired end reads with all of the NGS platforms considered here also facilitates the identification of genome rearrangements when relative mapping orientations or positions of reads do not correspond to those expected from the reference genome (e.g. [7]).

Short sequencing reads can also be used to identify potential nucleotide variations due to RNA editing, a post-transcriptional mechanism for which specific bases are substituted or inserted/deleted [64]. RNA editing by base substitution is the most frequent type of editing and has been well investigated in mitochondria of land plants [65]. RNA-Seq reads from Illumina and SOLiD platforms have been successfully used to detect the complete editing pattern in the mitochondrial genome of grapevine, supporting the idea that RNA editing in plant mitochondria is likely more pervasive than expected (Picardi *et al.*, submitted for publication). In mammals, several known editing events have been accurately detected using 454 sequencing technology [66].

Large-scale transcriptome analysis by RNA-Seq

NGS platforms are ideally suited for the detailed analysis of the transcriptome. Indeed, our current knowledge of the transcriptome complexity in different tissues, cell types, developmental stages and physiological or pathological conditions is very partial, even in widely studied organisms such as human or mouse. Alternative splicing, a pervasive phenomenon affecting in human virtually all multi-exon genes [67–69] is a major determinant of transcriptome complexity. Indeed, in human, a mean of at least ten different variants are observed for each gene—increasing the expression potential of the genome by at least an order of magnitude. The deep-sequencing coverage provided by NGS platforms is expected to revolutionize the detection and quantification of expressed transcripts. This novel technology, termed RNA-Seq, provides sequence reads from one (single-end sequencing) or both (paired-end sequencing) ends of cDNAs generated by a population of total or polyA enriched RNAs.

The nature of the sample is critical. Typically, RNA-Seq analyses are carried out on the poly-A enriched fraction to specifically detect protein coding mRNAs. However, in this way a functionally relevant part of the transcriptome, consisting of non-polyadenylated ncRNAs, can be missed. To obtain a more comprehensive overview of the transcriptome the random amplification of total RNA can be carried out, taking care to perform a rRNA depletion step to prevent an unwanted saturation of sequence reads from the rRNA fraction.

The read length, ranging from 30 bp for ABI SOLiD and Illumina to over 400 bp for Roche 454 FLX, with the corresponding level of throughput, defines the optimal range of applicability of the three different NGS platforms.

The huge throughput and short read length of Illumina and SOLiD (see Table 1) make these two technologies more suitable for quantifying transcript levels through tag profiling [8, 70] also termed digital gene expression, and full-length transcript profiling [71]. The latter methodology suitably applies to the detection of transcribed regions in the genome, refining known exon coordinates and discovering novel ones. However, as such reads typically span a single exon the relevant information about exon connectivity is missing and makes problematic the detection and relative quantification of expressed full length isoforms. For this aim the longer reads produced by Roche 454 FLX are much more informative although sophisticated model-based systems [72] are intended to deconvolute the relative abundance of different transcripts derived from the same gene.

RNA-Seq may also be very effective for the discovery of novel splice sites and splicing variants. The discovery of novel splice sites can be carried out either by searching contiguous mappings against splice junction libraries derived from the concatenation of all known 5′ and 3′ splice junctions [68]

or performing *de novo* splice site discovery by using mapping tools like QPALMA [73] or TopHat [74] specifically designed for *ab initio* detection of reads spanning exon junctions, thus able to perform split alignments against the reference genome.

It should be noted that whenever RNA-Seq reads span an exon boundary contiguous genome mapping strategies are not expected to find correct matches. For this reason it is advisable, when using a contiguous mapper (see Table 2) to carry out the mapping not only against the genome but also against the full set of known transcripts as derived from RefSeq [75] and other databases such as ASPicDB [76] that also collect alternative transcript variants not represented in RefSeq.

Another issue to be considered is the strand specificity of RNA-Seq data. cDNA sequences produced by Roche 454 FLX can be in either orientation and the discrimination of sense from antisense transcripts is not necessarily trivial. On the other hand SOLiD and more recently Illumina provide kits for obtaining strand specific sequences.

As previously anticipated RNA-Seq data may also allow the accurate quantification of transcript levels. Mortazavi *et al.* [67] proposed the Reads Per Kilobase of exon model per Million mapped measures (RPKM). RPKM is simply given by:

$$RPKM = 10^9 \times \frac{C}{N \cdot L},$$

where C is the number of mappable reads that fell onto the gene's exons, N is the total number of mappable reads in the experiment and L is the total length of the exons.

A new generation of bioinformatics tools attempt transcriptome annotation using only RNA-Seq data (e.g. [77]). However, several studies have shown that various steps in sample preparation (mRNA fragmentation, use of oligo-dT versus random primers for cDNA synthesis, size selection of fragments for sequencing, etc.) can introduce substantial biases into the distribution of reads along templates [78–81]. Such phenomena can impact upon many applications of NGS, but are particularly important for RNA-seq and can complicate both transcript annotation and quantification.

CHiP-Seq

Chromatin Immunoprecipitation (CHiP) [82] refers to the isolation of genomic fragments bound to proteins through the use of crosslinking agents and specific antibodies to identify genomic regions bound to histones or specifically by DNA binding proteins such as TFs. This technology is rapidly becoming the method of choice for the large-scale identification of TF–DNA interactions, or, more broadly, of the characterization of chromatin packaging—how genomic DNA is packaged into histones and in correspondence with which histone modifications. Chip-Seq implies the characterization of isolated DNA by NGS approaches (as opposed to the search for specific sequences by PCR, or the identification of isolated DNA through microarray-based approaches). Genomic fragments may be subjected to single or paired end sequencing strategies and reads are mapped to the genome to identify enriched regions—in principle those that contain functional binding sites for the factors of interest.

Once reads have been mapped to the reference sequence, it is necessary to determine which regions are flanked by a sufficient number of reads to discriminate them from 'background' noise due to sequence errors, contamination of isolated protein–DNA complexes, non-specific protein binding and other stochastic factors.

One way to filter out noise is to use a negative control to generate a pattern of noise to be compared to the read map generated from the real data (either using an antibody which does not recognize any TF, or by using a cell type that does not express the factor of interest). It is clear that genomic regions enriched only in the positive experiment should be those of interest.

In the absence of control experiments, background read levels must be estimated using stochastic methods. If we assume that in a completely random experiment each genomic region has the same probability of being extracted and sequenced, given t, the overall number of tags, and g, the size of the genome, then the probability of finding one tag mapping in a given position is given by t/g. The same idea can be applied by dividing the genome into separate regions (for example, the chromosomes or chromosome arms), since for experimental reasons different regions can have different propensities to produce reads. Thus, global or region-specific 'local' matching probability can be calculated, and the expected number of tags falling into any genomic region of defined size can be estimated for example using Poisson or negative binomial distributions. Finally, the significance of tag enrichment is computed, by using sliding windows across the whole

genome. If a 'control' experiment is available, the number of tags it produced from a given region can serve directly as 'background' model.

Several 'peak-finding' methods have been published, including FindPeaks [83], F-Seq [84], SISSRS [85], QuEST [86], MACS [87], the ChipSeq Peak Finder used in [88], ChIPDiff [89] and CisGenome [90], which encompasses a series of tools for the different steps of the ChIP-seq analysis pipeline. False discovery rates are estimated by these tools by comparing the level of enrichment (number of tags) at given sites, with the background model used.

The reliability of 'peak finding' methods has yet to be fully evaluated, since most of them were devised for a single experiment, and their portability to different organisms and/or experimental conditions is often not clear. Moreover, with current tools the choice of a significance or enrichment threshold to discriminate real binding sites from background is often not immediate and left to users, based on calculated false discovery rates and/or on the level of enrichment of the expected binding motif and/or prior knowledge about genomic regions bound by the TFs themselves. In fact, the first applications of ChIP-Seq have been related to epigenetic regulation (see for example [91, 92]), perhaps because the problem is somewhat easier than for TFs. The analysis protocol is the same, with the difference that TFs can bind DNA with different affinities resulting in 'grey areas' of tag enrichment, while the detection of histone modifications is more of a 'yes or no' decision, making the separation between signal and noise in peak detection much clearer. Primary analyses of CHiP-seq data for TF–DNA binding are thus invariably followed by further efforts to validate the predictions and to identify the short motifs bound by factors of interest at the genomic loci identified. Such efforts vary from ChIP-PCR, to the recognition of known motifs, to the detection of sequences overrepresented in isolated fragments. A detailed description of such strategies is beyond the scope of this review.

Small RNAs

Recent years have seen number of important discoveries relating to the regulated expression of small (typically 18–25 base) RNAs in eukaryotic cells and their important roles, principally as regulators of stability or availability for translation of mRNAs, with which they can interact by means of base complementarity e.g. [93] but also as guides for genome methylation [94] and potentially in other processes. Deep sequencing of small RNAs has become the method of choice for small RNA discovery and expression analysis [95]. Unlike oligonucleotide array studies, deep sequencing requires no a-priori knowledge of the nature of small RNAs, is less subject to the lack of specificity of short probes sometimes associated with oligonucleotide arrays [96] and expression levels can be followed over a wider range with deep sequencing. Indeed, even the shortest sequencing reads will yield the complete sequence of a 'small RNA', making these molecules ideal targets for characterization by NGS technologies. Given that typical sequencing runs will include parts of adaptors used in preparation of cDNA, it is critical that partial adaptor sequences are removed before analysis. This can be achieved through custom scripts, using the sequence file pre-processing tool from the UEA plant sRNA toolkit [97], tools from the Bioconductor open source package (http://bioconductor.org/) or using the 'vectorstrip' or 'fuzznuc' programs from the EMBOSS package [98]. It should also be born in mind that additional bases, not derived from the genome sequence are often added physiologically to the 3′ ends of mature microRNAs and these bases can also obscure correct alignments to genomic sequences [99].

Many classes of small RNAs exist as families present as multiple highly conserved copies within a single genome and often conserved between related organisms. Clustering of observed sequences and comparison with databases of annotated small RNAs (e.g. miRBase [100], piRNAbank [101] and the Arabidopsis Small RNA Project [102]) allows the identification of members of conserved families and provides indications as to their relative expression levels. Analysis of the size distribution of reads can also prove informative as to the nature of small RNAs present. For example, microRNAs tend to be ~21 bases in length as are the transactivating small RNAs tasi-RNAs) of plants, other siRNAs in plants typically being 24 bases in length while piRNAs of animals tend to be between 25 and 33 bases in length.

Several specific bioinformatics tools have been developed to identify members of different classes of small RNAs from deep sequencing data. Tools such as mirDeep [103] and MirCat [97] exploit structural characteristics of miRNA precursors by

mapping small RNA reads to the genome of origin and searching for plausible hairpin structures encompassing regions where small RNAs map and identifying cases where a single species, deriving from a stem is over-represented with respect to a putative miRNA* and reads derived from loop regions. For a review dedicated to the discovery and expression profiling of miRNA using deep sequencing, see ref. [104].

Transactivating siRNAs (ta-siRNAs) are a class of 21 base small interfering RNAs in plants that show derive from longer double stranded RNA precursors [105]. The characteristic phasing of ta-siRNAs derived from common precursors has been exploited to develop an algorithm to identify statistically significant phasing in alignments of small RNAs to genomic sequences. This algorithm has also been implemented as a web tool [97].

Piwi associated RNAs (piRNAs) are a class of repeat associated small interfering RNAs (ra-siRNAs) derived from large repeat containing genomic loci such as the flamenco locus in Drosophila, predominantly expressed in germline cells, and thought to be principally involved in the regulation of transposon expression through complementary interactions leading to degradation of transcripts [106]. piRNAs appear to use various members of the PIWI subfamily of argonaute proteins in a distinctive amplification loop mediated by reciprocal cleavage of piRNA precursors and target molecules [107]. A simple algorithm to detect such complementary patterns has been proposed [108].

Finally, several groups have recently proposed an elegant strategy exploiting the fact that, in plants, complementarity between miRNAs/ta-siRNAs and their targets usually leads to precise cleavage of target mRNAs [109,110]. These workers sequenced the 5′ ends of mRNA degradation products, assuming that sites targeted by siRNAs would be over-represented. A dedicated bioinformatics pipeline for matching end-reads, datasets of known small RNAs and a database of transcripts has been presented [111].

Epigenomics studies

5′-Methylation of cytosine bases forms the basis of important mechanisms of regulation of chromatin state and gene expression [112]. It is becoming increasingly clear that DNA methylation and demethylation can be a dynamic process in both animals [113] and plants [114]. One of the most popular methods of characterizing the methylation state of genomic DNA has been the targeted sequencing of particular genomic regions after treatment of isolated DNA with bisulfite which converts unmethylated cytosines to uracil, but does not modify 5′ methylated cytosines. More recently, and analogously to the situation with ChiP experiments, specifically designed microarrays have allowed the identification of methylated and non-methylated regions though hybridization with bisulfite treated genomic DNA. The development of NGS technologies has provided an alternative approach whereby bisulfite treated DNA is directly sequenced and mapping of reads to the genomic sequence allows identification of methylated sites and quantification of the frequency with which such sites are methylated DNA (for a comprehensive review, see [115]).

Clearly, modification of non-methylated cytosines will increase the level of mismatches in reads derived from non-methylated regions and potentially introduce artifactual matches to regions of the genome other than the one from which reads were derived. Amplification of genomic DNA fragments adds additional complications (antisense reads derived from modified or non-modified genomic regions). The mapping of bisulfite reads is relatively straightforward for Roche 454 data where conventional mapping tools can recover statistically significant matches. Of the limited numbers of studies of this type published until now using Illumina data, two have used conventional short read mapping tools and both native and computationally modified genome sequences (where for each strand methylation modifications have been performed) in order to identify reads that map uniquely to a single genome locus [116,117], while a third [87] developed a novel probabilistic mapping procedure based on base call scores and combinatorial substitution of cytosines for thymines in reads. To minimize the computational cost of exhaustive genome scans for each read, an efficient branch and bound algorithm was applied to an appropriate genome index structure, to exclude genomic regions that could not include significant matches. While the use of NGS technologies in epigenomic studies is in its infancy, the increasing awareness of the importance of epigenetic marking in development and disease suggest that this field will develop rapidly over the next years, at both the experimental and bioinformatics levels.

CONCLUSIONS

We have attempted to provide a broad outline of bioinformatics approaches for the analysis of NGS data. The rapid rate of development in the field means that it is likely that significant developments will have occurred even by the time of publication of this review. To this end we have avoided detailed discussions of data formats and quality scores. However, several dynamic and useful discussion forums on the WWW may be of use to keep up-to-date with recent developments (e.g. http://seqanswers.com, http://groups.google.com/group/solexa). Finally, it is fascinating to speculate as to what questions will next be addressed using these potent technologies. Metatranscriptomics for example [118–121], promises to allow previously unimaginable advances in our understanding of large scale biological interactions in microbial communities. In terms of 'conventional' genome sequencing, either current or novel NGS technologies will undoubtedly contribute to the goal of providing plausible 'personal genomics' services whereby complete genome sequences of individuals will contribute to improved diagnostics and therapeutic programming while requiring novel tools for data management and to ensure data privacy. In the meantime many informed observers believe that the goal of a $1000 human genome sequence [122] is almost within reach and several technologies are on the verge of meeting the X prize challenge of sequencing 100 human genomes in 10 days at a cost of no more than $10 000 per genome (http://genomics.xprize.org/). Indeed, NGS technologies are already playing a key role in the 1000 genomes project [123] directed at the wide sampling of human genome sequences. There cannot be any doubt that NGS approaches are here to stay and will provide major stimuli for bioinformatics for many years to come, both at the level of algorithm development and Laboratory Information Management System (LIMS) development and implementation (essential for the accurate management and archiving of the volumes of data generated in modern post-genomic research). Here we have focused on the current generation of bioinformatics tools for analysis of NGS data, which tend to be command line driven and somewhat inaccessible to many wet-bench researchers. There is undoubtedly a need for more intuitive, graphic user interface instruments to render the power of these new technologies available to a wider audience within the scientific community. All of these considerations will further enhance the symbiotic relationship between modern biology and computational sciences, and ensure long and productive careers for talented and committed bioinformaticians.

Key Points

- NGS technologies are revolutionizing the scale and perspectives of research in the fields of genomics and functional genomics.
- The general features of the three major NGS platforms, namely Roche 454, Illumina Solexa and AB SOLID, are illustrated.
- NGS data require 'next-generation bioinformatics' for the handling and the analysis of the huge amount of data produced.
- A simulation carried out by using two benchmarks datasets against the human genome and transcriptome illustrates current limitations and open problems in genome mapping of NGS data.
- The major bioinformatics applications for dealing with NGS including genome mapping, *de novo* assembly, detection of SNPs and editing sites, transcriptome analysis, ChIP-Seq, small RNA characterization and epigenomic studies are briefly discussed.

FUNDING

FISM (Associazione Italiana Sclerosi Multipla), AIRC (Associazione Italiana Ricerca sul Cancro), Telethon (grant number GGP06158), Progetto Strategico Regione Puglia PST_012, Ministero Università e Ricerca (progetto FIRB, laboratorio Internazionale di Bioinformatica) and Ministero Politiche Agricole e Forestali (VIGNA consortium).

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977; **74**:5463–7.
2. MacLean D, Jones JD, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 2009;**7**:287–96.
3. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet* 2008;**24**:133–41.
4. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;**9**:387–402.
5. Mardis ER. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med* 2009;**1**:40.
6. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 2008;**92**:255–64.
7. Morozova O, Marra MA. From cytogenetics to next-generation sequencing technologies: advances in the detection of genome rearrangements in tumors. *Biochem Cell Biol* 2008;**86**:81–91.

8. Morrissy AS, Morin RD, Delaney A, *et al*. Next-generation tag sequencing for cancer gene expression profiling. *Genome Res* 2009;**19**:1825–35.
9. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods* 2008;**5**:16–18.
10. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol* 2009;**25**:195–203.
11. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
12. Lister R, Gregory BD, Ecker JR. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr Opin Plant Biol* 2009;**12**:107–18.
13. Droege M, Hill B. The Genome Sequencer FLX System—longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol* 2008;**136**:3–10.
14. Bennett S. Solexa Ltd. *Pharmacogenomics* 2004;**5**:433–8.
15. Porreca GJ, Shendure J, Church GM. Polony DNA sequencing. *Curr Protoc Mol Biol* 2006;**Chapter 7:**Unit:7–8.
16. Harris TD, Buzby PR, Babcock H, *et al*. Single-molecule DNA sequencing of a viral genome. *Science* 2008;**320**:106–9.
17. Branton D, Deamer DW, Marziali A, *et al*. The potential and challenges of nanopore sequencing. *Nat Biotechnol* 2008; **26**:1146–53.
18. Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* 2008;**26**: 602–11.
19. Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. *Genomics* 2009;**93**:105–11.
20. Ewing B, Hillier L, Wendl MC, *et al*. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998;**8**:175–85.
21. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64.
22. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–410.
23. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. *Nat Biotechnol* 2009;**27**:455–7.
24. Bateman A, Quackenbush J. Bioinformatics for next generation sequencing. *Bioinformatics* 2009;**25**:429.
25. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 2008;**24**: 2395–6.
26. Lin H, Zhang Z, Zhang MQ, *et al*. ZOOM! Zillions of oligos mapped. *Bioinformatics* 2008;**24**:2431–7.
27. Li R, Li Y, Kristiansen K, *et al*. SOAP: short oligonucleotide alignment program. *Bioinformatics* 2008;**24**:713–14.
28. Campagna D, Albiero A, Bilardi A, *et al*. PASS: a program to align short sequences. *Bioinformatics* 2009;**25**:967–8.
29. Eaves HL, Gao Y. MOM: maximum oligonucleotide mapping. *Bioinformatics* 2009;**25**:969–70.
30. Abouelhoda MI, Kurtz S, Ohlebusch E. The enhanced suffix array and its applications to genome analysis. *Algorithms Bioinformatics Proc* 2002;**2452**:449–63.
31. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
32. Li R, Yu C, Li Y, *et al*. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 2009;**25**:1966–7.
33. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 2009;**25**: 1754–60.
34. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008;**18**:1851–8.
35. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* 2008;**9**:128.
36. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* 2009;**25**:1363–9.
37. Rumble SM, Lacroute P, Dalca AV, *et al*. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* 2009;**5**:e1000386.
38. Kuhn RM, Karolchik D, Zweig AS, *et al*. The UCSC genome browser database: update 2009. *Nucleic Acids Res* 2009;**37**:D755–61.
39. Richter DC, Ott F, Auch AF, *et al*. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One* 2008; **3**:e3373.
40. Aury JM, Cruaud C, Barbe V, *et al*. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* 2008;**9**:603.
41. Reinhardt JA, Baltrus DA, Nishimura MT, *et al*. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae pv. oryzae*, *Genome Res* 2009;**19**:294–305.
42. Petrosino JF, Highlander S, Luna RA, *et al*. Metagenomic pyrosequencing and microbial identification. *Clin Chem* 2009;**55**:856–66.
43. Havlak P, Chen R, Durbin KJ, *et al*. The Atlas genome assembly system. *Genome Res* 2004;**14**:721–32.
44. Batzoglou S, Jaffe DB, Stanley K, *et al*. ARACHNE: a whole-genome shotgun assembler. *Genome Res* 2002;**12**: 177–89.
45. Huang X, Wang J, Aluru S, *et al*. PCAP: a whole-genome assembly program. *Genome Res* 2003;**13**:2164–70.
46. Mullikin JC, Ning Z. The phusion assembler. *Genome Res* 2003;**13**:81–90.
47. Bryant DW, Jr, Wong WK, Mockler TC. QSRA: a quality-value guided de novo short read assembler. *BMC Bioinformatics* 2009;**10**:69.
48. Butler J, MacCallum I, Kleber M, *et al*. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* 2008;**18**:810–20.
49. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**:821–9.
50. Hernandez D, Francois P, Farinelli L, *et al*. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* 2008;**18**:802–9.
51. Jeck WR, Reinhardt JA, Baltrus DA, *et al*. Extending assembly of short DNA sequences to handle error. *Bioinformatics* 2007;**23**:2942–4.
52. Dohm JC, Lottaz C, Borodina T, *et al*. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res* 2007;**17**: 1697–706.
53. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. *Genome Res* 2008;**18**:324–30.

54. Warren RL, Sutton GG, Jones SJ, *et al*. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 2007; **23**:500–1.
55. de Bruijn NG. A combinatorial problem. *Koninklijke Nederlandse Akad v. Wetenschappen* 1946;**49**:758–64.
56. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinform* 2009;**10**:354–66.
57. Taillon-Miller P, Gu Z, Li Q, *et al*. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genoome Res* 1998;**8**:748–54.
58. Mooney S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform* 2005;**6**:44–56.
59. Marth GT, Korf I, Yandell MD, *et al*. A general approach to single-nucleotide polymorphism discovery. *Nat Genet* 1999; **23**:452–6.
60. Smith DR, Quinlan AR, Peckham HE, *et al*. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res* 2008;**18**:1638–42.
61. Ondov BD, Varadarajan A, Passalacqua KD, *et al*. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics* 2008;**24**:2776–7.
62. Malhis N, Butterfield YS, Ester M, *et al*. Slider—maximum use of probability information for alignment of short sequence reads and SNP detection. *Bioinformatics* 2009;**25**: 6–13.
63. Quinlan AR, Stewart DA, Stromberg MP, *et al*. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* 2008;**5**:179–81.
64. Gray MW. Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life* 2003;**55**:227–33.
65. Takenaka M, Verbitskiy D, van der Merwe JA, *et al*. The process of RNA editing in plant mitochondria. *Mitochondrion* 2008;**8**:35–46.
66. Wahlstedt H, Daniel C, Enstero M, *et al*. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res* 2009;**19**: 978–86.
67. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
68. Pan Q, Shai O, Lee LJ, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;**40**: 1413–15.
69. Wang ET, Sandberg R, Luo S, *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008; **456**:470–6.
70. Valen E, Pascarella G, Chalk A, *et al*. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res* 2009;**19**:255–65.
71. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009; **10**:57–63.
72. Zheng S, Chen L. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Res* 2009;**37**:e75.
73. De Bona F, Ossowski S, Schneeberger K, *et al*. Optimal spliced alignments of short sequence reads. *Bioinformatics* 2008;**24**:i174–80.
74. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**: 1105–11.
75. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;**35**:D61–5.
76. Castrignano T, D'Antonio M, Anselmo A, *et al*. ASPicDB: a database resource for alternative splicing analysis. *Bioinformatics* 2008;**24**:1300–4.
77. Denoeud F, Aury JM, Da Silva C, *et al*. Annotating genomes with massive-scale RNA sequencing. *Genome Biol* 2008;**9**:R175.
78. Dohm JC, Lottaz C, Borodina T, *et al*. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008;**36**:e105.
79. Harismendy O, Frazer K. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. *Biotechniques* 2009;**46**:229–31.
80. Harismendy O, Ng PC, Strausberg RL, *et al*. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 2009;**10**:R32.
81. Quail MA, Kozarewa I, Smith F, *et al*. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 2008;**5**:1005–10.
82. Collas P, Dahl JA. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front Biosci* 2008; **13**:929–43.
83. Fejes AP, Robertson G, Bilenky M, *et al*. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* 2008;**24**:1729–30.
84. Boyle AP, Guinney J, Crawford GE, *et al*. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* 2008;**24**:2537–8.
85. Jothi R, Cuddapah S, Barski A, *et al*. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 2008;**36**:5221–31.
86. Valouev A, Johnson DS, Sundquist A, *et al*. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* 2008;**5**:829–34.
87. Cokus SJ, Feng S, Zhang X, *et al*. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 2008;**452**:215–19.
88. Johnson DS, Mortazavi A, Myers RM, *et al*. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007; **316**:1497–502.
89. Xu H, Wei CL, Lin F, *et al*. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* 2008;**24**: 2344–9.
90. Ji H, Jiang H, Ma W, *et al*. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 2008;**26**:1293–300.
91. Barski A, Cuddapah S, Cui K, *et al*. High-resolution profiling of histone methylations in the human genome. *Cell* 2007;**129**:823–37.
92. Mikkelsen TS, Ku M, Jaffe DB, *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007;**448**:553–60.

93. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**:281–97.
94. Baulcombe D. RNA silencing in plants. *Nature* 2004;**431**: 356–63.
95. Meyers BC, Axtell MJ, Bartel B, *et al*. Criteria for annotation of plant MicroRNAs. *Plant Cell* 2008;**20**:3186–90.
96. Barad O, Meiri E, Avniel A, *et al*. MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. *Genome Res* 2004;**14**:2486–94.
97. Moxon S, Schwach F, Dalmay T, *et al*. A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* 2008; **24**:2252–3.
98. Olson SA. EMBOSS opens up sequence analysis. European molecular biology open software suite. *Brief Bioinform* 2002; **3**:87–91.
99. Morin RD, O'Connor MD, Griffith M, *et al*. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* 2008; **18**:610–21.
100. Griffiths-Jones S, Grocock RJ, van Dongen S, *et al*. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006;**34**:D140–4.
101. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res* 2008;**36**:D173–7.
102. Gustafson AM, Allen E, Givan S, *et al*. ASRP: the Arabidopsis Small RNA Project Database. *Nucleic Acids Res* 2005;**33**:D637–40.
103. Friedlander MR, Chen W, Adamidi C, *et al*. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 2008;**26**:407–15.
104. Creighton CJ, Reid JG, Gunaratne PH. Expression profiling of microRNAs by deep sequencing. *Brief Bioinform* 2009;**10**:490–7.
105. Vazquez F, Vaucheret H, Rajagopalan R, *et al*. Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Mol Cell* 2004;**16**:69–79.
106. Brennecke J, Aravin AA, Stark A, *et al*. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* 2007;**128**:1089–103.
107. Gunawardane LS, Saito K, Nishida KM, *et al*. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in Drosophila. *Science* 2007;**315**:1587–90.
108. Olson AJ, Brennecke J, Aravin AA, *et al*. Analysis of large-scale sequencing of small RNAs. *Pac Symp Biocomput* 2008; **13**:126–36.
109. Addo-Quaye C, Eshoo TW, Bartel DP, *et al*. Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr Biol* 2008;**18**:758–62.
110. German MA, Pillay M, Jeong DH, *et al*. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol* 2008;**26**:941–6.
111. Addo-Quaye C, Miller W, Axtell MJ. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 2009;**25**:130–1.
112. Weber M, Schubeler D. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr Opin Cell Biol* 2007;**19**:273–80.
113. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 2008;**9**:465–76.
114. Henderson IR, Jacobsen SE. Epigenetic inheritance in plants. *Nature* 2007;**447**:418–24.
115. Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res* 2009;**19**: 959–66.
116. Lister R, O'Malley RC, Tonti-Filippini J, *et al*. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008;**133**:523–36.
117. Meissner A, Mikkelsen TS, Gu H, *et al*. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;**454**:766–70.
118. Gilbert JA, Field D, Huang Y, *et al*. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* 2008;**3**: e3042.
119. Poretsky RS, Gifford S, Rinta-Kanto J, *et al*. Analyzing gene expression from marine microbial communities using environmental transcriptomics. *J Vis Exp* 2009, Feb 18(24) pii:1086. doi:10.3791/1086.
120. Shi Y, Tyson GW, DeLong EF. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* 2009;**459**:266–9.
121. Warnecke F, Hess M. A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *J Biotechnol* 2009;**142**:91–5.
122. Service RF. Gene sequencing. The race for the $1000 genome. *Science* 2006;**311**:1544–6.
123. Siva N. 1000 Genomes project. *Nat Biotechnol* 2008;**26**: 256.