BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 1 2011, pages 9–13
doi:10.1093/bioinformatics/btq603

Genome analysis

Advance Access publication October 22, 2010

GAMES identifies and annotates mutations in next-generation sequencing projects

Maria Elena Sana[1], Maria Iascone[2], Daniela Marchetti[2], Jeff Palatini[3], Marco Galasso[1] and Stefano Volinia[1,4,*]

[1]DAMA, Data Mining for Analysis of DNA, Department of Morphology and Embryology and TecnoPolo for Life Sciences, University of Ferrara, 44121 Ferrara, [2]Genetica Molecolare, Ospedali Riuniti di Bergamo, 24128 Bergamo, Italy, [3]Comprehensive Cancer Center, Ohio State University, Columbus, 43210 OH and [4]Molecular Virology Immunology & Medical Genetics and Biomedical Informatics, Ohio State University, Columbus, OH, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Next-generation sequencing (NGS) methods have the potential for changing the landscape of biomedical science, but at the same time pose several problems in analysis and interpretation. Currently, there are many commercial and public software packages that analyze NGS data. However, the limitations of these applications include output which is insufficiently annotated and of difficult functional comprehension to end users.

Results: We developed GAMES (Genomic Analysis of Mutations Extracted by Sequencing), a pipeline aiming to serve as an efficient middleman between data deluge and investigators. GAMES attains multiple levels of filtering and annotation, such as aligning the reads to a reference genome, performing quality control and mutational analysis, integrating results with genome annotations and sorting each mismatch/deletion according to a range of parameters. Variations are matched to known polymorphisms. The prediction of functional mutations is achieved by using different approaches. Overall GAMES enables an effective complexity reduction in large-scale DNA-sequencing projects.

Availability: GAMES is available free of charge to academic users and may be obtained from http://aqua.unife.it/GAMES.

Contact: stefano.volinia@unife.it

Received on August 11, 2010; revised and accepted on October 14, 2010

1 INTRODUCTION

Next-generation sequencing (NGS), or deep sequencing, methods are rapidly changing the standards in genomics (Metzker, 2010). Consolidated instruments, such as Illumina Genome Analyzer (Illumina Inc., San Diego, CA, USA) and ABI SOLiD (Applied Biosystems, Foster City, CA, USA) can sequence a full human genome in previously unimagined speed. To date, NGS has been applied in various contexts, including whole-genome sequencing (Bentley, 2006), discovery of transcription factor binding sites, and non-coding RNA expression profiles (Kato, 2009; Mardis, 2008). Massively parallel DNA-sequencing technologies, combined with sequence-capture methodologies (targeted re-sequencing) have obvious potential in clinical diagnostics, particularly for complex disorders that may be caused by a combination of several genes. A major goal of these studies is the detection of single nucleotide polymorphisms (SNPs), insertions and deletions (InDels), and other rearrangements to identify disease-associated variants in clinical samples. These genetic determinants contribute to variation in phenotypes, risk to diseases, and response to drugs or to the environment. Different studies, such as 1000 Genomes Project (Kaiser, 2008) and the Cancer Genome Atlas (http://cancergenome.nih.gov/), have implemented massive sequencing to more efficiently catalog genetic mutations responsible for cancer and other diseases. Therefore, NGS provides important knowledge about genetic variants in the population and about those genetic variants that are being commonly used in diagnostic studies in clinical settings. On one hand, NGS provides unprecedented opportunities for high-throughput genetic research; alternatively, it poses problems in rationale data analysis and requires adequate computational strategies to organize, handle and interpret the results (Shendure and Ji, 2008).

In the past few years, tools for the analysis of short/long reads datasets have become available. Currently, the scientific community has access to many commercial and open source software packages for analyzing NGS data. CLC Genomics Workbench software (CLC bio, Aarhus, Denmark) and NextGENe (SoftGenetics, State College, PA) are among the most popular in the commercial domain. Furthermore, there are open source packages, such as MAQ (Li *et al.*, 2008), SHRIMP (Rumble *et al.*, 2009), PASS (Campagna *et al.*, 2009), BWA (Li and Durbin, 2009), SAMTOOLS (Li *et al.*, 2009), BFAST (Homer *et al.*, 2009), PerM (Chen *et al.*, 2009), SNVMix (Goya *et al.*, 2010 ; Shah *et al.*, 2009), Crossbow (Langmead *et al.*, 2009) and Atlas-SNP2 (Shen *et al.*, 2010).

For 1000 Genome Project pilot data bioinformatic analysis, researchers used Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik), MAQ, Coronalite (Applied Biosystems, Foster City, CA, USA), and SSAHA2 (Ning *et al.*, 2001) as aligners, and used Genome Analysis Toolkit (GATK) genotyper (McKenna *et al.*, 2010), PolyScan (Chen *et al.*, 2007) and PolyBayes (Marth *et al.*, 1999) for SNP detection.

These tools process data from various platforms using different algorithms (such as Bayesian approaches) and SNP detection procedures. However, their output often seems to be insufficiently annotated (to the detriment of clinical implications) and/or of difficult functional interpretation. To solve this problem, we

*To whom correspondence should be addressed.

developed a pipeline for high-level elucidation of NGS results. In this article, we introduce GAMES (Genomic Analysis of Mutations Extracted by Sequencing), a new tool for mining functional mutations. The main purpose of GAMES is to supply life science investigators with a detailed biological insight into genetic events. GAMES provides maximum flexibility and allows high-throughput analysis of data from different NGS platforms. GAMES generates documents for logical viewing and seamless inspection of alignments and mismatches. Ultimately, GAMES provides an easy-to-use framework for processing NGS data. It enables users to leverage the throughput and accuracy of analysis, while facilitating its translation into biomedical meaningful results.

2 METHODS

GAMES is written in PERL 5.8 using BioPerl (Stajich *et al*., 2002). The workflow in GAMES is composed of two sequential steps (Fig. 1). The first step generates per-position information in the pileup format; the second step processes, evaluates and annotates the identified genetic events.

2.1 Input files

The input of GAMES is the alignment in SAM (Sequence Alignment/Map) format (Li *et al.*, 2009). This format supports short and long reads and is compatible with different sequencing platforms. Several aligning packages, such as BWA, MAQ, BFAST and PerM can generate SAM alignments. To optimize performance, we used Binary Alignment/Map (BAM) format, the binary representation of SAM. The input, sorted by genomic coordinates, is processed by removing duplicate reads, in order to minimize artifacts and false positives. The final position-sorted BAM file is indexed (to create a bai file, a standard part in the SAM specification) and used as input to GAMES.

2.2 SNP/SNV/InDel calling and Pileup

To begin the analysis, GAMES requires a reference genome sequence in fasta format and the coordinate-sorted BAM file.

GAMES analyzes the alignment and, for each mismatch, performs SNP/SNV/InDel calling and extracts the position and the respective base in the reference sequence. During this step, GAMES displays the alignment in pileup format. Each output line consists of chromosome, position, reference base, consensus base (the first best call), Phred-like consensus quality, number of hits of reads covering the position, the second best call and the respective quality score and counts. The last two columns are relative to total counts and repetitivity. There are several parameters that the user can set for particular purposes and each configuration can alter speed and behavior of the script. The user sets mapping quality, reads' length (according to the experiment), minimal quality threshold, minimum coverage, minor allele frequency and maximal repetitivity. The parameters used to define the goodness of mutation detection are the mapping quality, a measure of the confidence that a read is correctly aligned (Li *et al*., 2008) and the Phred's base-specific quality score (Ewing and Green, 1998; Ewing *et al*., 1998). This score is logarithmically linked to error probabilities. The default minimum quality score is 20, corresponding to 99% of base-call accuracy. In the calculation of quality parameter, GAMES accounts for the minimum consensus quality in the three flanking nucleotides on each side of the mismatch. Considering the quality values in the neighboring bases improves the reliability at polymorphism detection (Altshuler *et al*., 2000). The minimum coverage is an integer calculated at each genomic position and indicates the minimum number of alignments covering the position. The minor allele frequency is the lowest frequency that can be tolerated in a heterozygote. Finally, in an optional procedure, GAMES blats each DNA segment (of read length) containing a bona fide genetic event to the reference genome (Kent, 2002) and calculates the repetitivity, defined as the number of loci to which the segment can be uniquely mapped. Non-unique elements can be filtered out. Together with mapping quality filter, this function assures a finer removal of repetitive sequences that could severely affect mapping accuracy.

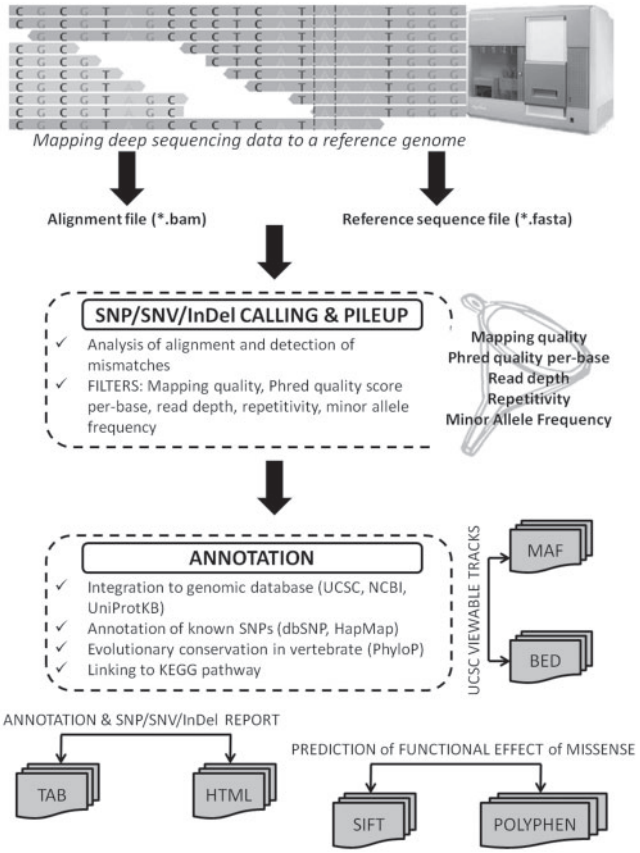

Fig. 1. The basic workflow of GAMES. Starting from alignment in BAM format to a reference genome, GAMES performs SNP/SNV/InDel calling and displays alignment in pileup format. Then it annotates the mismatches according to a host of databases (UCSC, dbSNP and HapMap). GAMES produces also viewable tracks (MAF and BED files) and outputs ready for SIFT and POLYPHEN-2, to predict the structural effect of mutations.

2.3 Annotation

GAMES has been developed primarily for both a comprehensible in-depth investigation and an accurate statistical selection of genetic variants. Building on data acquired from the pileup, GAMES extracts information about each mismatch and provides genome annotation by integration with genomic databases. GAMES uses UCSC Genome Browser (Kent *et al*., 2002), which provides a MySQL database annotation data. GAMES queries various tables in the UCSC database to extract the information for the reference genome (Table 1). From the *knownGene* table, we obtain the information relative to the genomic location of the nucleotide, the gene, the chromosome, the exon count (for all available isoforms), the strand and the encoded protein/s. We use this information to determine the position of the mismatch in the gene and whether it is in a coding/non-coding, exon, intron, UTR or intron/exon junction region (intronic regions contiguous to exon starts and exon ends that are important to evaluate splice-site mutations). To find

Table 1. SQL tables used by GAMES to provide genomic annotation of NGS in human

UCSC *table*	*Information*
knownToRefSeq	UCSC Gene ID → RefGene ID
knownGene	Gene symbol, strand, txStart, txEnd, cdsStart, cdsEnd, exonCount, exonStarts, exonEnds, proteinID
phyloPwayAll	Conservation phyloP score
snp130	dbSNP ID, strand, observed, class, valid
hapmapSnpsCEU	HapMap ID, strand, observed, allele1, homoCount1, allele2, homoCount2, heteroCount
keggPathway	KEGG pathway map ID
keggMapDesc	KEGG pathway map description

the correspondence between RefGene ID and UCSC Gene ID, we employ the table *knownToRefSeq*. Once the script has filtered the mismatch, or a different genetic event, and its relative genomic location, it searches for a correspondence in databases of known SNPs: dbSNP (Smigielski *et al*., 2000) and HapMap (Frazer *et al*., 2003). In the current release of GAMES, we uses *snp130* (build 130) and *hapmapSnpsCeu* tables. The *snp130* table contains almost 19 million SNP annotations, and includes the first set of SNP calls from the 1000 Genomes Project. The *hapmapSnpsCeu* table consists of a set of approximately four million common SNPs, and the genotyping of these SNPs in different human populations. The tables are structured to include the position of the SNP on the genome assembly as well as additional informations, such as sequences of the observed alleles from rs-fasta files, genotype counts and allele frequencies, kind of mutation and the pathogenetic significance of the SNPs, if reported. GAMES queries the correspondence between the known SNP in dbSNP and the event identified by the pileup analysis. If any correspondence exists, GAMES marks this event as 'known SNP', otherwise, it lists it as 'unknown SNP'. If the genomic location is not reported in dbSNP, it is indicated as 'unreported'. From HapMap database, GAMES extracts the allele frequencies reported in the population.

GAMES also annotates genetic events with the measurement of evolutionary conservation using the PhyloP score for 44 vertebrates (Siepel *et al*., 2005, 2006). In the PhyloP table, sites predicted to be conserved are assigned positive scores, while sites predicted to be fast evolving are assigned negative scores. Functional mutations are expected to be associated with conserved sites, while polymorphisms are expected to be associated with variable locations. Additional information that GAMES extracts from databases relates to the pathways in which the gene is involved. From *keggPathway* and *keggMapDesc* tables, GAMES annotates what is the metabolic or pathologic pathway for each gene through the interface with KEGG (Kanehisa and Goto, 2000; Kanehisa *et al*., 2006, 2010). This type of information is critical for the proper analysis of whole-exome sequencing projects, in which mutations in different genes might be biologically involved in the same pathway or disease. GAMES incorporates information about the distribution of reads in the forward and reverse strands. This information is necessary because the cross-correlation between the forward and reverse tag counts can have significant effects on the accuracy of the mutation detection (Bansal, 2010). An unbalanced distribution of reads between forward and reverse strands may be indicative of sequencing artifacts and, consequently, may lead to false discoveries. This parameter can then be used to further filter bona fide genetic events.

2.4 Output files

All results generated by GAMES are deposited in text and html files. These files are distributed by a private web server, where they are cataloged according to the respective NGS project. A text file lists all the genetic events detected and their relative annotations. This file is in tab-delimited columns format. Starting on the left side of the page, the columns report genomic information, such as the gene symbol, RefGene ID and UCSC Gene ID (GAMES takes into account all known isoforms in each gene), the UniprotKB ID for the associated protein/s, the exon count, the chromosome and the coordinates, the strand of the gene and the coding sequence limits. GAMES differentiates coding or non-coding regions, exons, introns or intron/exon junctions. The columns, following toward right, report the first base call and the second base call with the respective quality score, count and frequency. If needed, GAMES calculates the effect of the mutation in the primary sequence of the protein. The links to dbSNP and HapMap report known SNPs and the eventual clinical relevance. In the last group of columns, GAMES lists the repetitivity of the local sequence (if the BLAT option is enabled), the nucleotide conservation in vertebrates and the pathway in which the gene is involved (with the KEGG ID). The most useful and convenient results are displayed as interactive tables, in an html file, that enlists only those mutations covering the coding regions and the intron/exon junctions. In this table, the items are directly linked to main databases: UCSC Genome Browser, Entrez Gene (NCBI), UniprotKB, dbSNP (NCBI), HapMap (NCBI), Variation Viewer (NCBI) for clinically associated SNPs. The genetic events are sorted according to the different parameters chosen by the investigator. The prediction of the structural effect of a mutation is very important to have a clinical outline of the disease. To allow this analysis at protein level, GAMES lists all missense mutations in coding region, not reported in dbSNP or HapMap databases, in two different text files that can be directly used by POLYPHEN-2 (Adzhubei *et al*., 2010) and SIFT (Kumar *et al*., 2009). These tools predict whether an amino acid substitution affects protein function, respectively, by a probabilistic classifier based on machine-learning method and by sequence homology and the physical properties of amino acids. Additionally, GAMES generates a BED and a MAF track for each sample, viewable on graphical interfaces. They can be visualized with Broad's Integrative Genomics Viewer (IGV) (http://www.broadinstitute.org/igv), EnsEMBL or UCSC Genome Browser. The BED (Browser Extensible Display) file contains four fields: the chromosome, the starting and the ending position of the feature in the chromosome, the coverage of the site in the reads. The MAF (Multiple Alignment Format) file stores the multiple alignment between the reads and the genome, underlying the mismatches in the sequence. A composite MAF file can include the alignment between the reference genome and different samples to allow comparative genomes variation parsing. Finally, an index is generated in the web site folder that summarizes and links to all the different reports for that sample.

3 RESULTS AND DISCUSSION

Initially, we validated GAMES by using reads obtained with Genome Analyzer (Illumina) after SureSelect Target Enrichment of 36 genes involved in hereditary hypertrophic cardiomyopathy (Wheeler *et al*., 2009). The selected regions included 1.9 Mb of DNA from different human chromosomes (NCBI build 36). We sequenced eight samples from patients with hereditary hypertrophic cardiomyopathy.

We used BWA (Burrows–Wheeler Alignment) to align 35 bp single-end reads to the human genome. BWA builds assemblies by mapping short reads to reference sequences performing a heuristic Smith–Waterman-like alignment. The SAM file of the alignment was converted to BAM file using SAMTOOLS. Preprocessing of the Illumina-sequences dataset was performed to remove PCR duplicates and trim those sequences that contained bases with low quality scores (threshold = 20) .

GAMES processed the alignment to detect about 2200 mismatches between the human genome reference and each sample, with minimum quality score of 20 (corresponding to an accuracy of 99%) and minimum read depth of 10. The minor allele frequency was

set to 25% (less common variant at a biallelic SNP). The observed mismatch detection frequency in these eight samples was thus on average of 1.1 mismatches/kilobase (2200 mismatches/1900 kb). These results were in excellent agreement with the SNP density reported by the HapMap consortium (1.14 genotyped polymorphic SNPs per kilobase). The mismatches covering coding regions were about 130 and of these roughly 30 were unknown SNPs in each sample.

All a priori known SNPs, detected in the samples by applying standard Sanger sequencing, were also obtained by NGS and confirmed in the GAMES analysis. Additionally, four-fifths of the novel SNPs detected by NGS/GAMES were experimentally validated by Sanger sequencing (data not shown).

GAMES was also used for the analysis of mutation segregation, and parental transmission of mismatches in a whole-exome project. This technique selectively analyzes the 1% of the genome that contains genes coding for proteins.

Additionally, we tested GAMES using 2×100 bp paired-end reads produced with Genome Analyzer IIx (Illumina) and obtained essentially comparable results with the single-end runs.

Finally, GAMES was also applied to SOLiD data ($25 + 50$ bp paired end) from chronic leukemia samples. The results restated the flexibility and the accuracy of this pipeline for the mining, sorting and annotation of genetic variation from different NGS platforms.

4 CONCLUSION

We have developed and validated GAMES, a new application for the processing and the annotation of NGS data. GAMES generates concise and highly readable reports for a functional selection of important genetic events. GAMES is a flexible tool, it can be used for the analysis of data from different NGS platforms, it supports short and long, single-end and paired-end reads. In fact, GAMES uses as input the alignment file in the standard generic SAM format, which is obtainable from many free and commercial alignment tools.

The specificity and the accuracy of this application are warranted by the use of a range of independent filters, which are geared toward the selection of statistically and functionally significant genetic events.

The investigator can easily and promptly use GAMES reports to discriminate silent mutations or polymorphisms from variants that are potentially associated with a phenotype or a disease. Moreover, by using BED and MAF files as tracks in UCSC Genome Browser or other viewers, it is possible to swiftly identify mutations in different samples or individuals and to interrogate their biological relevance. SIFT and POLYPHEN-2 outputs can be used to predict the effect of the mutations in proteins.

In conclusion, GAMES is not only a new application for mining functional SNPs or InDels from NGS data, but it also aids in the overall interpretation of large-scale sequencing data. GAMES' primary purpose is to provide a tool for life science investigators, who do not have bioinformatic expertise, to obtain biological insight into genetic events.

ACKNOWLEDGEMENTS

We thank Pearlly Yan and Hansjuerg Alder at OSUMC for helpful discussion. We also thank Susan Lutz for manuscript revision.

Funding: AIRC (Italian Association for Cancer Research) and BioPharmaNet (Rete Alta Tecnologia dell'Emilia Romagna) grants to S.V.

Conflict of Interest: none declared.

REFERENCES

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Altshuler,D. *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Bansal,V. (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, **26**, i318–i324.

Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.

Campagna,D. *et al.* (2009) PASS: a program to align short sequences. *Bioinformatics*, **25**, 967–968.

Chen,K. *et al.* (2007) PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Res.*, **17**, 659–666.

Chen,Y. *et al.* (2009) PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*, **25**, 2514–2521.

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

Ewing,B. *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.

Frazer,K.A. *et al.* (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Goya,R. *et al.* (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.

Homer,N. *et al.* (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, **4**, e7767.

Kaiser,J. (2008) DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science*, **319**, 395.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **28**, 27–30.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Kato,K. (2009) Impact of the next generation DNA sequencers. *Int. J. Clin. Exp. Med.*, **2**, 193–202.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

Langmead,B. *et al.* (2009) Searching for SNPs with cloud computing. *Genome Biol.*, **10**, R134.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Marth,G.T. *et al.* (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, **23**, 452–456.

McKenna,A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Ning,Z. *et al.* (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.

Rumble,S.M. *et al.* (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.*, **5**, e1000386.

Shah,S.P. *et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.

Shen,Y. *et al.* (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.*, **20**, 273–280.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Siepel,A. *et al.* (2006) New methods for detecting lineage-specific selection. *Res. Comput. Mol. Biol. Proc.*, **3909**, 190–205.

Smigielski,E.M. *et al.* (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.

Stajich,J.E. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Wheeler,M. *et al.* (2009) A new era in clinical genetic testing for hypertrophic cardiomyopathy. *J. Cardiovasc. Transl. Res.*, **2**, 381–391.