APPLICATIONS OF NEXT-GENERATION SEQUENCING — OPINION

Next-generation sequencing data interpretation: enhancing reproducibility and accessibility

Anton Nekrutenko and James Taylor

Abstract | Areas of life sciences research that were previously distant from each other in ideology, analysis practices and toolkits, such as microbial ecology and personalized medicine, have all embraced techniques that rely on next-generation sequencing instruments. Yet the capacity to generate the data greatly outpaces our ability to analyse it. Existing sequencing technologies are more mature and accessible than the methodologies that are available for individual researchers to move, store, analyse and present data in a fashion that is transparent and reproducible. Here we discuss currently pressing issues with analysis, interpretation, reproducibility and accessibility of these data, and we present promising solutions and venture into potential future developments.

Today, one often hears that life sciences are faced with the 'big data problem'. However, data are just a small facet of a much bigger challenge. The true difficulty is that most biomedical researchers have no capacity to carry out analyses of modern data sets using appropriate tools and computational infrastructure in a way that can be fully understood and reused by others. This struggle began with the introduction of microarray technology, which, for the first time, introduced life sciences to truly large amounts of data and the need for quantitative training[1–3]. What is new, however, is that next-generation sequencing (NGS) has made this problem vastly more challenging. Today's sequencing-based experiments generate substantially more data and are more broadly applicable than microarray technology, allowing for various novel functional assays, including quantification of protein–DNA binding or histone modifications (using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq)[4]), transcript levels (using RNA sequencing (RNA-seq)[5]), spatial interactions (using Hi-C[6]) and others. These individual applications can be combined into larger studies, such as the recently published genomic profiling of a human individual whose genome was sequenced and gene expression tracked over an extended period in a series of RNA-seq experiments[7]. As a result, meaningful interpretation of sequencing data has become particularly important. Yet such interpretation relies heavily on complex computation — a new and unfamiliar domain to many of our biomedical colleagues — which, unlike data generation, is not universally accessible to everyone.

Here we make the case that any future progress within the life sciences acutely depends on the democratization of biomedical computation so that even the smallest research units with modest budgets are capable of carrying out analyses using appropriate tools in a reproducible fashion. Making such democratization possible involves many layers discussed here, including: developing best practices; removing obstacles associated with using heterogenous software on complex high-performance computing infrastructure; facilitating the interactive exploration of analysis parameters; and, perhaps most importantly, promoting the concepts of analysis transparency and reproducibility. To give the reader a sense of immediate urgency, we survey a number of recent studies that use NGS technologies and that show the lack of general agreement on how data analyses are to be carried out. We specifically highlight the fact that very few current studies record exact details of their computational experiments, making it difficult for others to repeat them.

Adoption of existing analysis practices

As mentioned above, there are numerous applications of NGS technologies. Yet there are common analysis challenges among all of these applications. Here we use one type of NGS application — variant discovery — as an example. In this analysis, which is becoming common in medical genetics and serves as the foundation for future personalized medicine, genomic DNA is sequenced, and the resulting data are compared against a reference sequence to catalogue differences: such differences can range from SNPs to complex chromosomal rearrangements. A series of accepted practices for variant discovery is starting to emerge owing to efforts such as the 1000 Genomes Project[8] (also see BOX 1). One would expect that these approaches will be widely used in studies that feature a similar design. As we demonstrate below, this is not the case and is thus a cause for grave concern because not following tested practices undermines the quality of biomedical research, limiting its potential.

Owing to highly coordinated efforts, such as HapMap[9] and the 1000 Genomes Project[8], variant discovery has become one of the more developed areas of NGS data analysis, in which every procedure can be carried out with freely accessible open-source software (albeit even in variant discovery there is no 'single best' analysis strategy; see REF. 10 for an in-depth Review). It is therefore surprising that few publications follow this approach (BOX 1). Although most publications discussed in BOX 1 (all are resequencing studies with experimental design similar to that of the 1000 Genomes Project) used recommended mappers (13) and variant callers (12), only four studies used the complete workflow. So what prevents wide adoption of existing practices? A typical variant-calling analysis involves removing PCR duplicates, recalibrating quality scores and refining alignment, followed by genotyping, refining calls and annotating obtained variants[10]. For human data, a workflow using Picard and SAMtools[11], for sorting and indexing, and GATK[12], for recalibration, realignment and variant discovery, is well documented[13] and is used by the 1000 Genomes community. GATK requires that "all data sets (reads, alignments, quality scores, variants, dbSNP information, gene tracks, interval lists — everything) must be sorted in order of one of the canonical references sequences", such as the standard human genome version used by the 1000 Genomes Project. This is a necessary and unavoidable way of simplifying and optimizing the performance of downstream analyses. However, this also implies that all data sets provided as inputs to GATK must be resorted in the required order and supplied with read-group information. Although GATK can be used for analysis of non-human data[14], it is difficult to apply this toolkit to non-model organisms, as numerous steps of the analysis (for example, score recalibration and realignment) rely on auxiliary data sets that are available for only a handful of well-annotated genomes. Data in BOX 1 suggest that these logistical challenges may be too complex for most researchers, who choose to use more straightforward approaches, potentially sacrificing the quality of their results.

The difficulty of reproducibility

The procedure recommended by the 1000 Genomes Project is certainly not the only way to carry out genotyping. In fact, alternative approaches used in publications listed in BOX 1 may yield comparable results. Additionally, the entire field of NGS analysis is in constant flux, and there is little agreement on what is considered to be the 'best practice'. In this situation, it is especially important to be able to reuse and to adopt various analytical approaches reported in the literature. Unfortunately, this is often difficult owing to the lack of necessary details. Let us look at the first and most straightforward of the analyses: read mapping. To repeat a mapping experiment, it is necessary to have access to primary data and to know the software and its version, parameter settings and name of the reference genome build. From the 19 papers listed in BOX 1 and in Supplementary information S1 (table), only six satisfy all of these criteria. To investigate this further, we surveyed 50 papers (BOX 2) that use the Burrows–Wheeler Aligner (BWA)[15] for mapping (the BWA is one of the most popular mappers for Illumina data). More than half do not provide primary data and list neither the version nor the parameters used and neither do they list the exact version of the genomic reference sequence. If these numbers are representative, then most results reported in today's publications using NGS data cannot be accurately verified, reproduced, adopted or used to educate others, creating an alarming reproducibility crisis.

Box 1 | 1000 Genomes Project as an example of best practices and their adoption

The 1000 Genomes Project is an international effort aimed at uncovering human genetic variation with the ultimate goal of understanding the complex relationship between genotype and phenotype. The project aims to uncover all genetic variants with a frequency of at least 1% in major population groups from Europe, East Asia, South Asia, West Africa and the Americas. The project consists of the two phases: the pilot phase and the main phase. The objective of the pilot phase was to evaluate technologies as well as to develop and to fine-tune the methodological framework for the analysis of data sets in the main phase. The pilot phase consisted of three experiments that used distinct strategies, including low-coverage (2–4×) whole-genome sequencing of 179 unrelated individuals (pilot 1), high-coverage (20–60×) sequencing of two father-mother-child trios (pilot 2) and high-coverage (50×) sequencing of 1,000 targeted gene regions in 900 individuals (pilot 3). The main phase involves sequencing approximately ~2,500 individuals from 27 populations at low coverage (~4×).

The results of the pilot phase were published in 2010 (REF. 8) (also see REF. 34 for a description of data and access; the analysis of the main phase is currently underway). The crucial importance of the pilot project was the establishment of a series of best practices that are used for the analysis of the data from the main phase and are broadly applicable for analysis of resequencing data in general. These are described at the 1000 Genomes analysis description and software tools pages and were discussed in detail in a 2011 Review in this journal[10].

Despite the fact that analytical procedures developed by the project are well documented and rely on freely accessible open-source software tools, the community still uses a mix of heterogeneous approaches for polymorphism detection: we surveyed 299 articles published in 2011 that explicitly cite the 1000 Genomes pilot publication[8]. Nineteen of these were in fact resequencing studies with an experimental design similar to that of the 1000 Genomes Consortium (that is, sequencing a number of individuals for polymorphism discovery; Supplementary information S1 (table)). Only ten studies used tools recommended by the consortium for mapping and variant discovery, and just four studies used the full workflow involving realignment and quality score recalibration. Interestingly, three of the four studies were co-authored by at least one member of the 1000 Genomes Consortium.

We note that this discussion thus far has dealt only with technical reproducibility challenges or with the ability to repeat published analyses using the original data to verify the results. Most biomedical researchers are much more acquainted with biological reproducibility, in which conceptual results are verified by an alternative analysis of different samples. However, we argue that the computational nature of modern biology blurs the distinction between technical and biological reproducibility. Consider the following example. Numerous cancer resequencing studies have been published in 2011, including two in patients with head and neck squamous cell carcinoma (HNSCC). The first study carried out exome resequencing in 32 patients and identified 17 missense mutations in *NOTCH1* in 21 individuals, suggesting a novel function for this locus in the aetiology of HNSCC[16]. The other study published in the same issue carried out resequencing in an independent set of 92 patients and identified seven *NOTCH1* missense mutations in 11% of studied tumours[17]. There was no overlap between the two sets of missense mutations identified in the two studies. To understand why the two studies arrived at different sets of *NOTCH1* alterations (which may well be a legitimate outcome), it is necessary to evaluate the technical aspects of each study that require sufficient analysis details and primary data — a lack of these would prevent such an evaluation.

The potential of integrative resources

The above challenges can be distilled into two main issues. First, most biomedical researchers experience great difficulty carrying out computationally complex analyses on large data sets. Second, there is a lack of mechanism for documenting analytical steps in detail. Recently, a number of solutions have been developed that, if widely adopted, would solve the bulk of these challenges. These solutions can collectively be called integrative frameworks, as they bring together diverse tools under the umbrella of a unified interface. These include BioExtract[18], Galaxy[19], GenePattern[20], GeneProf[21], Mobyle[22] and others (see REF. 21). These tools record all analysis metadata, including tools, versions and parameter settings used, ultimately transforming research provenance for a task that requires active tracking by the analyst to one that is completely automatic. Additionally, these resources enable biomedical researchers seamlessly to make use of powerful computing infrastructure, which is necessary for the analysis of NGS data (for example, the personal 'omics' data generated by REF. 7 are close to 0.5 terabytes in size). We elaborate on these two points below.

Making life sciences transparent and reproducible. The overwhelming majority of currently published papers using NGS technologies include analyses that are not detailed (BOXES 1,2). Moreover, the computational approaches used in these publications cannot be readily reused by others. Integrative frameworks described above provide the ideal medium for tracking, recording and disseminating all details of computational analyses. For example, GenePattern pioneered automatic embedding of analysis details into Microsoft Word documents while preparing publications[23]. The Galaxy system's Pages function can be used to create interactive Web-based supplements for research publications with the analysis details (data sets, histories and workflows) directly embedded. These documents are then published as standard Web pages that can be viewed in any modern browser, allowing readers to inspect the described analysis down to the finest levels of detail (see REF. 24 for an example).

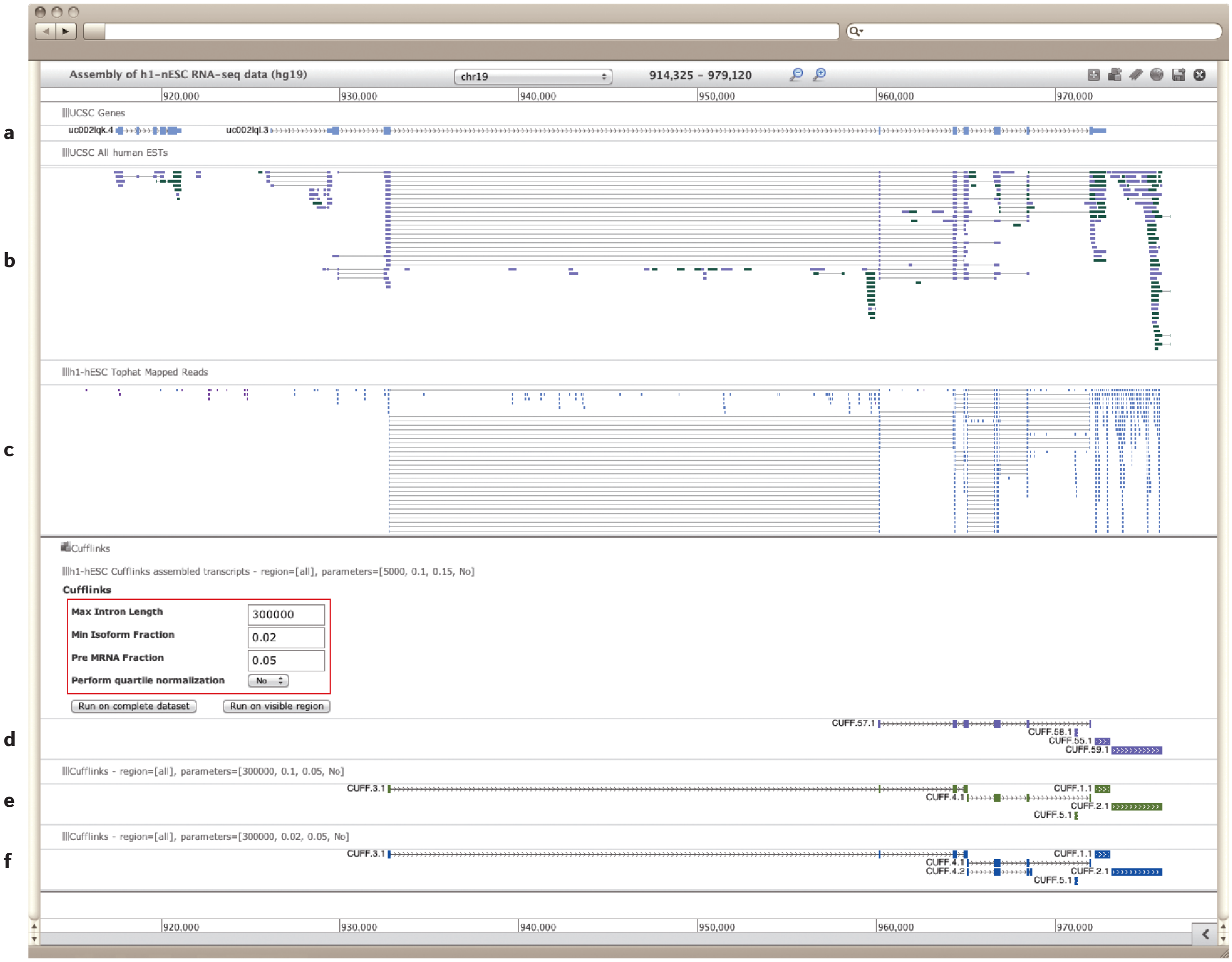
Making high-performance computing infrastructure useful to biologists. Analysis of NGS is computationally demanding. There are numerous examples of excellent high-performance computing (HPC) resources that can be used for NGS computation. These include large computing clusters available at numerous institutions and nationwide efforts such as XSEDE, as well as private and public clouds (for example, Amazon Elastic Compute Cloud). These resources combine large numbers of processors with extensive storage and fast network interconnects, but they are of little use to the broad biomedical community because special skills are often required to harness the full power of these resources. For example, substantial informatics expertise is required because cloud resources are presented to users either as virtual machines or as application-programming interfaces. Several commercial vendors now provide 'software as a service' analysis solutions for sequence data (such as DNAnexus and GenomeQuest); however, these solutions are built on proprietary platforms, raising concerns about transparency and vendor lock-in. Others have built open-source software packages that can scale well on cloud resources, but these are typically single-purpose (such as Crossbow for variant discovery[25], Myrna for RNA-seq analyses[26] and CloVR for metagenomic and small-genome annotation projects[27]) and thus may not be easily integrated into larger workflows.

Integrative frameworks make complexity and heterogeneity of HPC resources transparent to biologists. Anyone should be able to deploy an integrative solution on any type of resource, whether cloud, cluster, desktop or otherwise. One example of such a deployment model that is specifically targeted at biomedical research is CloudMan[28,29], which is an extensible framework and user interface for managing computing clusters on cloud resources. CloudMan allows users to create computing clusters on the fly without directly interacting with cloud providers and to use such clusters as a computational back end for integrative resources, such as Galaxy or any other system (CloudMan has been used with Galaxy but is not specifically coupled to it).

Improving long-term archiving. One important problem associated with making analyses reproducible is the longevity of hosted analysis services. For example, the Galaxy page that describes the analysis of mitochondrial heteroplasmy[24] is not guaranteed to be online forever. This highlights a general vulnerability of centralized resources: their very existence can depend on various external factors, such as a changing funding climate. A promising solution is to offer the users of integrative frameworks the ability to create snapshots of a particular analysis. For example, cloud computing vendors generally provide the ability to take snapshots of a virtual machine, including stored data. Sharing such a snapshot provides an excellent approach for transparent communication and reproducibility. A link to a virtual machine could be included with a publication, and reviewers or readers could start their own exact copy of that machine to inspect and to verify analyses. However, there are also issues with this as an approach for long-term archiving. Typically, the author must continue to pay the cloud provider to maintain the snapshot, or it will no longer be available. Integrative frameworks resolve this situation by providing support for composing virtual machine images from a specific analysis that can be stored as an archival resource such as the Dryad system or Figshare. By tracking connections between data sets, analysis workflows and results, these frameworks could allow users to export a complete collection of analysis results automatically for archival by

Box 2 | Barriers to reproducibility are widespread

Many classical publications in life sciences have become influential because they provide complete information on how to repeat reported analyses so others can adopt these approaches in their own research, such as for chain termination sequencing technology that was developed by Sanger and colleagues[35] and for PCR[36,37]. Today's publications that include computational analyses are very different. Next-generation sequencing (NGS) technologies are undoubtedly as transformative as DNA sequencing and PCR were more than 30 years ago. As more and more researchers use high-throughput sequencing in their research, they consult other publications for examples of how to carry out computational analyses. Unfortunately, they often find that the extensive informatics component that is required to analyse NGS data makes it much more difficult to repeat studies published today. Note that the lax standards of computational reproducibility are not unique to life sciences; the importance of being able to repeat computational experiments was first brought up in geosciences[38] and became relevant in life sciences following the establishment of microarray technology and high-throughput sequencing[3,39,40]. Replication of computational experiments requires access to input data sets, source code or binaries of exact versions of software used to carry out the initial analysis (this includes all helper scripts that are used to convert formats, groom data, and so on) and knowing all parameter settings exactly as they were used. In our experience (BOX 1 and Supplementary information S1 (table)), publications rarely provide such a level of detail, making biomedical computational analyses almost irreproducible. Supplementary information S2 (reference list) lists 50 papers randomly selected from 378 manuscripts published in 2011 that use the Burrows–Wheeler Aligner[15] for mapping Illumina reads. Most papers (31) provide neither a version nor the parameters used, and neither do they provide the exact version of the genomic reference sequence. From the remaining 19 publications, only four studies provide settings, eight studies list the version, and only seven studies list all necessary details. More than half of the studies (26 out of 50) do not provide access to the primary data sets. In two cases, authors provided links to their own websites, where data were deposited; however, in both cases, links were broken.


Assembly of h1-nESC RNA-seq data (hg19)
chr19
914,325 – 979,120
920,000
930,000
940,000
950,000
960,000
970,000
a
UCSC Genes
uc002lqk.4
uc002lql.3
UCSC All human ESTs
b
h1-hESC Tophat Mapped Reads
c
Cufflinks
h1-hESC Cufflinks assembled transcripts - region=[all], parameters=[5000, 0.1, 0.15, No]
d
CUFF.57.1
CUFF.58.1
CUFF.55.1
CUFF.59.1
Cufflinks - region=[all], parameters=[300000, 0.1, 0.05, No]
e
CUFF.3.1
CUFF.4.1
CUFF.1.1
CUFF.2.1
CUFF.5.1
Cufflinks - region=[all], parameters=[300000, 0.02, 0.05, No]
f
CUFF.3.1
CUFF.4.1
CUFF.4.2
CUFF.1.1
CUFF.2.1
CUFF.5.1

Cufflinks

Max Intron Length	300000
Min Isoform Fraction	0.02
Pre MRNA Fraction	0.05
Perform quartile normalization	No

Run on complete dataset | Run on visible region

Figure 1 | **A prototype visual analytics framework for next-generation sequencing analysis.** Consider a researcher who is attempting to reconstruct mouse transcripts from RNA sequencing (RNA-seq) data using the TopHat and Cufflinks set of tools[45]. The quality of transcripts assembled in this analysis depends to a large extent on the choice of parameters, resulting in the procedure being repeated and initiating a wasteful cycle of tweaking parameters and re-running the analysis many times. If the visualization is instead tightly coupled with the analysis in an approach called visual analytics, the researcher can instead modify parameters and observe a graphical representation of resulting transcript assembly changes in real time. Here, a region of mouse chromosome 9 is being visualized in Galaxy's dynamic interactive 'Trackster' browser. Tracks shown from the top are: gene annotations (**a**) and expressed sequence tag (EST) sequences (**b**) extracted from the UCSC Genome Browser; RNA-seq reads mapped using TopHat spliced alignment (**c**); and Cufflinks transcript assemblies produced using various parameters being set interactively (**d–f**). Note the inputs surrounded by the red box that correspond to four Cufflinks parameter settings: maximum intron length, minimum isoform fraction, precursor mRNA fraction and minimum SAM mapping quality. Changing parameters directly within the browser will change the structure of assembled transcripts, providing the visual feedback for parameter selection.

a journal or by another data repository. Using this archive, anyone will be able to recreate a new virtual instance with the exact tools, data and workflows associated with the archived analysis. This approach towards reproducibility avoids the need for a centralized resource providing archival of analysis and achieves an unprecedented level of reproducibility.

Looking ahead

Tool development. Today, most analysis tools are distributed through mechanisms such as SourceForge, GitHub, Google Code and others, including direct downloads from developers' websites. These mechanisms have limited appeal for biomedical researchers as software still needs to be compiled, installed and supplied with associated data (for example, genome indices required by short-read mappers). There must to be a better way to distribute biomedical software. Digital distribution platforms such as Apple App Store, as well as subsequent efforts by Amazon, Google, RIM and others, provide an example of a system in which applications are seamlessly delivered and installed on users' devices. We argue that implementing a similar system for biomedical software will revolutionize the field by providing biologists with access to a multitude of tools and analysis workflows. The first such efforts include GenePattern's GParc and the Galaxy Tool Shed. These systems are designed to store tools wrapped for a given integrative framework (in this case, either GenePattern or Galaxy). The core idea is simplicity: users select a set of tools that will then be automatically installed into the user's analysis environment instance. Ultimately, these systems must encompass more than just tools: they should allow sharing of analysis workflows, data sets, visualizations and any other analysis artefacts created within integrative frameworks.

Success of this distribution model will crucially depend on buy-in from individual tool developers. Whereas systems such as the Apple App Store are tightly controlled and distribute applications tailored towards a single operating system that runs on a well-defined set of hardware platforms, the NGS analysis software ecosystem is chaotic and is likely to stay this way (see REFS 30,31). In addition, the success of the commercial AppStore concept is due to the potentially great financial payoff to the developer in cases in which an application becomes popular. Yet even in the commercial App Store, many programs remain free, as developers see exposure to large numbers of users as the ultimate reward. Such exposure could be the incentive to attract developers of biomedical software because integrative frameworks have the potential to expose bioinformatic tools to large audiences of users, enabling tools to be more widely used and more able to compete for funding and recognition.

Integrating analysis and visualization for easier data interpretation. Visual assessment is the culmination of every genomic experiment. At present, genomic data visualization builds on the concept of a genome browser pioneered by Artemis[32], popularized by UCSC Genome Browser[33] and further extended by dozens of visualization frameworks that are currently in active use or development. Yet despite this remarkable progress, visualization in the genomic context is usually the last step of an analysis. This does not need to be the case, as making visualization an active component of NGS analyses will have radical implications. First, integrating analysis and visualization aids the analysis process by allowing the user to see how parameter changes affect the final result in real time (FIG. 1). Second, in the context of scientific publications, the visualization is likely when the reader starts to evaluate the study. The disconnection between the visual representation of data and the underlying analysis makes this evaluation difficult. Tight integration between visualizations, analyses and data will transform readers' abilities to evaluate and to inspect results that rely on complex computational analysis approaches.

Box 3 | Guidelines for reproducibility

- Accept that the computational component is becoming an integral component of biomedical research. As the life sciences are becoming increasingly data-driven, there will be no escape from computation and data handling. Familiarize yourself with best practices of scientific computing using existing educational resources, such as the Software Carpentry project[41]. Implementing good computational practices in your group will automatically take care of many of the points listed below.
- Always provide access to primary data. It is obvious that without access to the original data sets, any claims made in a publication cannot be verified. In situations in which the data cannot be made public (for example, clinical data sets under Institutional Review Board protection), they should be deposited in controlled access repositories (such as dbGaP[42]), where they can be retrieved by authorized users. One potential issue with this point is the fact that there is currently a debate on what constitutes primary data. Storing images generated by some next-generation sequencing (NGS) machines on a large scale has long been unfeasible. Public sequencing archives, such as those at the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI), are still accepting sequencing reads as submissions and should be used. Going forward, other formats, such as aligned data in BAM format, are likely to be used (as is already done by the 1000 Genomes Project).
- Record versions of all auxiliary data sets used during the analysis. For example, in most NGS analyses, such as variant discovery detailed here, sequencing reads are compared against a reference genome. It is crucial to record which reference genome was used because, just as software has versions and cars have model years, genomes have build identifiers. For example, the latest human genome build distributed by the UCSC Genome Browser is called hg19 (it is derived from the GRC37 build prepared by the Genome Reference Consortium) and has the highest number of functional annotations (7,330 annotation types) and should be the preferred version to be used. Note that the latest version may not always be the best choice. The latest mouse genome build (mm10) has only a fraction of annotations (258 tracks) compared with its predecessor (mm9, which has 2,096 tracks). Thus, it would be easier to interpret results of an NGS experiment mapped to the mm9 build even though mm10 has an additional 48 megabases of actual sequence.
- Note the exact versions of software used. Different versions of the same software often produce different results, and important bug fixes may have implications to results produced with a particular version of a tool.
- Record all parameters, even if defaults are used. Although the reason to record all parameters requires no explanation, we emphasize the importance of explaining default settings for reproducibility. A clause 'software was used with default settings' is found in many publications. However, the meaning of default settings often changes between versions of software and can be quite difficult to track down when a substantial amount of time has passed since publication. Thus, record what the default settings actually are.
- Provide all custom scripts. With the complexity of NGS analysis, it is often unavoidable to create simple scripts that carry out such straightforward tasks as, for example, changing data formats. Such scripts must be made accessible as any other part of the analysis.
- Do not reinvent the wheel. It pays to reuse existing software. Integrative frameworks and associated application stores already house hundreds of tools (for example, as of May 2012, Galaxy ToolShed contains ~1,700 tools). It is likely that a script for a particular problem has been already written. Ask around through existing resources such as SEQanswers[43] and BioStar[44].

Conclusions

Sustaining the growing application of NGS in biomedical research will require that data interpretation becomes as accessible as data generation. Moving biomedical research forwards will necessitate bridging the gap between experimentalists and computational scientists and adopting new practices (BOX 3). Experimentalists must now embrace unavoidable computational components of their research projects as being on a par with experimental activities in their laboratories. This would require including quantitative

Glossary

Application-programming interfaces

These define how different software components interact with each other. In the context of cloud computing, an application-programming interface defines how user-provided software interacts with the underlying cloud platform resources.

Virtual machines

A computing resource that appears to be a physical machine with a defined computing environment but may be simulated on another computing platform. In the context of cloud computing, virtual machines can be provisioned on demand and then accessed over the Internet like any other machine.

training in curricula, making it a requirement for attaining graduate degrees in life sciences and making sure that the details of computational experiments are recorded with the same level of care and scrupulousness as those of experimental procedures. At the same time, our computational colleagues must ask themselves if it is really possible for biologists to use their software. The emergence of integrative frameworks for accessible and reproducible analysis is a good indicator that things are starting to change, as the next big change in life sciences will come not from the new ways to generate data but from the innovative ways to analyse them. These frameworks enable us to follow the most important best practice in computational analysis: tracking details precisely and ensuring transparency and reproducibility.

Anton Nekrutenko is at the Huck Institutes of the Life Sciences and the Department of Biochemistry and Molecular Biology, Penn State University, Wartik 505, University Park, Pennsylvania 16802, USA.

James Taylor is at the Departments of Biology and Mathematics and Computer Science, Emory University, 1510 Clifton Road NE, Room 2006, Atlanta, Georgia 30322, USA.

e-mails: anton@bx.psu.edu; james.taylor@emory.edu

doi:10.1038/nrg3305

1. Allison, D., Cui, X. & Page, G. Microarray data analysis: from disarray to consolidation and consensus. *Nature Rev. Genet.* **7**, 55–65 (2006).
2. Quackenbush, J. Computational analysis of microarray data. *Nature Rev. Genet.* **2**, 418–427 (2001).
3. Ioannidis, J. P. A. *et al.* Repeatability of published microarray gene expression analyses. *Nature Genet.* **41**, 149–155 (2009).
4. Pepke, S., Wold, B. & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nature Methods* **6**, S22–S32 (2009).
5. Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
6. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
7. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293–1307 (2012).
8. Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
9. Gibbs, R., Belmont, J., Hardenbol, P. & Willis, T. The International HapMap Project. *Nature* **426**, 789–796 (2003).
10. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature* **12**, 443–451 (2011).
11. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
12. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
13. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
14. Auton, A. *et al.* A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**, 193–198 (2012).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
16. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in *NOTCH1*. *Science* **333**, 1154–1157 (2011).
17. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
18. Lushbough, C. An overview of the bioextract server: a distributed, web-based system for genomic analysis. *Adv. Comp. Biol.* **680**, 361–369 (2010).
19. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy Team Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
20. Reich, M., Liefeld, T., Gould, J., Lerner, J. & Tamayo, P. GenePattern 2.0. *Nature Genet.* **38**, 500–501 (2006).
21. Halbritter, F., Vaidya, H. J. & Tomlinson, S. R. GeneProf: analysis of high-throughput sequencing experiments. *Nature Methods* **9**, 7–8 (2011).
22. Néron, B., Ménager, H., Maufrais, C. & Joly, N. Mobyle: a new full web bioinformatics framework. *Bioinformatics* **25**, 3005–3011 (2009).
23. Mesirov, J. P. Accessible reproducible research. *Science* **327**, 415–416 (2010).
24. Goto, H. *et al.* Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol.* **12**, R59 (2011).
25. Langmead, B., Schatz, M., Lin, J. & Pop, M. Searching for SNPs with cloud computing. *Genome Biol.* **25**, 3005–3011 (2009).
26. Langmead, B. & Hansen, K. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.* **11**, R83 (2010).
27. Angiuoli, S. V. *et al.* CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* **12**, 356 (2011).
28. Afgan, E. *et al.* Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics* **11**, (Suppl. 12), S4 (2010).
29. Afgan, E. *et al.* Harnessing cloud computing with Galaxy Cloud. *Nature Biotech.* **29**, 972–974 (2011).
30. Stein, L. Creating a bioinformatics nation. *Nature* **417**, 119–120 (2002).
31. States, D. J. Bioinformatics code must enforce citation. *Nature* **417**, 588 (2002).
32. Parkhill, J., Crook, J., Horsnell, T. & Rice, P. Artemis: sequence visualization and annotation **16**, 944–945 (2000).
33. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
34. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nature Methods* **9**, 459–462 (2012).
35. Sanger, F. & Nicklen, S. DNA sequencing with chain-terminating inhibitors. *Bioinformatics* **24**, 104–108 (1977).
36. Saiki, R. *et al.* Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).
37. Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).
38. Schwab, M., Karrenbach, N. & Claerbout, J. Making scientific computations reproducible. *Comput. Sci. Engineer.* **2**, 61–67 (2000).
39. Carey, V. J. & Stodden, V. in *Biomedical Informatics for Cancer Research* (eds Ochs, M. F. *et al.*) 149–175 (2010).
40. Begley, C. G. & Ellis, L. M. Drug development: raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
41. Perkel, J. M. Coding your way out of a problem. *Nature Methods* **8**, 541–543 (2011).
42. Mailman, M., Feolo, M., Jin, Y., Kimura, M. & Tryka, K. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet.* **39**, 1181–1186 (2007).
43. Li, J., Schmieder, R., Ward, R. & Delenick, J. SEQanswers: an open access community for collaboratively decoding genomes. *Bioinformatics* **28**, 1272–1273 (2012).
44. Mangan, M., Miller, C. & Albert, I. BioStar: an online question & answer resource for the bioinformatics community. *PLoS Comp. Biol.* **7**, e1002216 (2011).
45. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protoc.* **7**, 562–578 (2011).

Acknowledgements

The authors are grateful for the support of the Galaxy Team (E. Afgan, G. Ananda, D. Baker, D. Blankenberg, D. Bouvier, D. Clements, N. Coraor, C. Eberhard, J. Goecks, J. Jackson, G. Von Kuster, R. Lazarus, R. Marenco and S. McManus). The visual analytics framework shown in Figure 1 has been built by J. Goecks. The authors' laboratories are supported by US National Institutes of Health grants HG005133, HG004909 and HG006620 and US National Science Foundation grant DBI 0850103. Additional funding is provided, in part, by the Huck Institutes for the Life Sciences at Penn State, the Institute for Cyberscience at Penn State and a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Anton Nekrutenko's homepage: http://www.galaxyproject.org
1000 Genomes Project: http://www.1000genomes.org
1000 Genomes analysis description: http://www.1000genomes.org/analysis
1000 Genomes software tools: http://www.1000genomes.org/tools
BioExtract: http://www.bioextract.org/GuestLogin
CloudMan: http://usecloudman.org
CloVR: http://clovr.org
Crossbow: http://bowtie-bio.sourceforge.net/crossbow/index.shtml
dbGaP: http://www.ncbi.nlm.nih.gov/gap
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP
Dryad: http://www.datadryad.org
Figshare: http://figshare.com
Galaxy: http://usegalaxy.org
Galaxy Tool Shed: http://usegalaxy.org/toolshed
GenePattern: http://www.broadinstitute.org/cancer/software/genepattern
GeneProf: http://www.broadinstitute.org/cancer/software/genepattern
GParc: http://gparc.org
HapMap: http://hapmap.ncbi.nlm.nih.gov
Mobyle: http://mobyle.pasteur.fr
Myrna: http://bowtie-bio.sourceforge.net/myrna/index.shtml
Picard: http://picard.sourceforge.net
SAMtools: http://samtools.sourceforge.net
Software Carpentry: http://software-carpentry.org
XSEDE: http://www.xsede.org

SUPPLEMENTARY INFORMATION

See online article: S1 (table) | S2 (reference list)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF