

Science in Focus – Bioinformatics Part 1 – Lost in Translation

Scope: 1500 words, ~ 30 references, 1 figure

Introduction

Bioinformatics was first coined in 1970 to refer to information stored in biological systems. But in today's terms it has become synonymous with the generation and interrogation of large scale biological data. The catalyst for this change in meaning was undoubtedly the completion of the Human Genome Project in the late 1990's [1]. It is instructive to remember that the primary goal of bioinformatics is to increase the understanding of biological processes, making it a fundamentally scientific discipline. What sets it apart is the use of intensive computational techniques to achieve this aim. This broad goal is perhaps one of the reasons that so many related areas are interrogated with a bioinformatics approach, everything from pure biological processes like genome assembly, protein structure analysis and RNA expression analysis, to data and text mining, to image analysis. Thus, it is an essential component of translational medicine.

From an oncological perspective the allure of bioinformatics has been its promise to unlock the wealth of the increasingly complex biological data generated from cancers. Cancer is fundamentally a genomic disease therefore one would assume if we can analyse the cancer's genome, its derivatives (RNA, proteins, metabolites), structure (chromatin, ploidy status) and regulation (methylation) we should be able to leverage this information to a therapeutic advantage. Indeed the ultimate expression of this aspiration is 'personalised medicine'; that every patient and their cancer can be broken down into its constituent parts and an individualised solution delivered.

As oncology enters 'big data', the conceptual framework of the dimensional challenges of 'big data' [2] has been neatly repurposed to help one understand the scale of this challenge [3]. Specifically, one needs to consider: 1) *Volume* – the amount of data generated, 2) *Variety* – the differing data sources from which we gather, including genomics but also electronic records, imaging, digital pathology, 3) *Velocity* – the rate at which data is generated by technological advances and the need to analyse such data in clinically relevant timescales, 4) *Value* to the clinician and the patient. Key steps essential to all 4 'V' are data collection, integration, interpretation and reproducibility, highlighting the central role of Bioinformatics in the 'big data' area.

Data collection

The quality of tissue entered into an analysis pipeline will tremendously influence the quality and reliability of the output data. Standard practice in all clinical pathology is to fix tissue in formalin and embed it in paraffin blocks (FFPE). Whilst this process has been integral to the generation of high quality pathology reporting, the process has lasting repercussions on the integrity of DNA [4,5] and RNA for multi-omic testing. Recently, alternative protocols have been proposed, with improvement on the results [6]. Fresh do not suffer the same artefacts producing higher quality data but such samples are not collected as standard practice. Tumour mutations are typically assessed using a targeted panel of known driver genes, for example APC, p53 and RAS in colorectal cancer. These are chosen empirically based on known biological processes in any given tumour site. The advantage of these approaches is their accessibility, the limited amount of data produced and the ability to sequence regions to high depth of coverage thereby increasing confidence in the final call [7]. However, the inherent bias of choosing the gene panel up front cannot be understated. Next generation sequencing (NGS) on the other hand provides an agnostic appraisal of either the whole genome, the whole exome or the transcriptome [8]. This allows for assessment of novel or less frequent mutations that are present at an individual level as well as variations in copy number and gene rearrangements. The trade-off however is that the depth of coverage is usually to a much lower level, thereby decreasing the confidence with a mutation can be 'called'. Neither approach is infallible but it is incumbent on those attempting to understand the clinical implications of the data to understand these technical limitations. The increasing availability and decreasing cost of NGS means that it is likely to become much more commonplace in medical applications [9,10]. Another important issue is that the sensitivity of a mutation call is influenced by the proportion of tumour contained within the clinical sample. Techniques such as microdissection will increase the yield of tumour but are not scalable and time consuming, and have the potential expense of understanding the associated stromal interactions with that tumour, overlooking valuable information on the tumour microenvironment. One emerging answer to this problem could be single cell sequencing which has seen high profile publications regarding the cellular heterogeneity of DNA, RNA, proteins and metabolites [11]. It has been argued that this sudden interest is due to three factors: technological advances that allow whole genome/transcriptome amplification, lower cost for higher throughput and the invention of technology for single cell manipulation [12]. Again we see examples of *volume* and *velocity*.

However, volume and speed do not necessarily guarantee a 'fair' assessment, which must rely on carefully planning and understanding of all potential confounders at the collection level. What type of tissue was analysed, on what platform and how best to integrate and analyse the information.

Integration and interpretation challenges

Being able to take a macroscopic overview of the volume of data generated, even when organised efficiently, is a monumental task. What complementary methods help to provide internal validation and different biological information, for example immunohistochemistry to validate the loss of heterozygosity in copy number outputs from NGS or to confirm transcriptomic effects with protein stains? Orthogonal practices such as digital analysis of the biopsy or resection site can greatly help but equally raises questions about which modality represents 'truth' and how to do you reconcile conflicting findings. A combination of both low and high throughput technologies (*variety*), and by applying novel bioinformatic and machine learning algorithms to existing data such as imaging and pathology slides [13–15], will allow maximal yield (*value*) from patient data.

For the data to be clinically useful it must be interpretable, which in turn means it has to be presented in a synthesised, succinct, informative fashion. The tasks involved with data integration and interpretation are usually the last steps in the data production pipeline but they are far from trivial, and have the power to influence the future of medical research and practice. Many examples could be provided here. We consider amongst them the SHIVA trial [16], a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. The aim was to assess the efficacy of molecularly targeted agents, chosen on the basis of tumour molecular profiling but used outside their indications, in patients with advanced cancer. This trial was ultimately unsuccessful in its primary aim, in that the use of molecularly targeted agents in a histology agnostic way, did not improve progression-free survival compared with treatment at physician's choice. More importantly, it started a heated discussion in the clinical and research communities when published in 2015[17]. The bioinformatic resourcing and effort to create a database to process and present all the data collected, within clinical trial timelines, has been commended [3]. That the trial was unsuccessful demonstrates that even with excellent data collection the interpretation and use of such data is far from straight forward. This includes understanding the role of co-activating mutations, cross talk between pathways at a transcriptomic level and the disease type itself, which was specifically ignored in SHIVA. The best example of this is BRAF V600E which is an excellent therapeutic target in melanoma [18] but not so in colorectal cancer [19]. Given how important these studies are for our understanding of disease and the implications they have on shaping our future clinical strategies, it is key that these efforts are extremely carefully designed, using input from a community of

scientists and medical experts ranging from tumour biologists, immunologists, geneticists, statisticians as well as the traditional clinical multidisciplinary team members.

Reproducibility challenges

Lack of reproducibility is a significant problem in biomedicine, and will further worsen if careful steps are not taken. Recently, frameworks have been developed, such as RIPOSTE [20], to encourage researchers to address fundamental bioinformatics and statistical issues at each stage of the process, right from the study design stage. Some of the requirements will be familiar to biomedical scientists and clinicians, however many will not be aware of the necessary quality control procedures for sample handling, data verification and cross validation to maximize reproducibility. For 'omic' studies, attempts must be made to control for known and unknown biases by verified processes of normalisation and batch effect correction. Hence, there is a real need to engage with bioinformaticians as early as possible in the study workflow, to ensure the correct steps are taken from data collection right to analysis and interpretation. Analysis scripts and algorithms for data processing and presentation should be readily available in repositories for dissemination and replication, as necessity[21,22]

The scale of 'big data' means that clinicians must also realise that there is a high potential false discovery rate (FDR) and adjustments must be made to account for multiple testing. Only up front discussions about data collection and an analysis plan can allow for minimisation of bias. The goal is therefore to embed this process into clinical trial design where so many 'knowns' are already accounted for. However, the velocity with which data accumulates exponentially outstrips to the rate at which trials get opened and completed. Therefore tissue bio-banking and data storage must be considered up front as much as is possible [23] with appropriate patient consent to allow for testing at later date.

Conclusions

Although patient journeys through health systems are linear in a temporal sense, there is a need to anticipate future requirements and think in a circular fashion about the processes that needs to take place in order to maximise the value of available data. As Figure 1 demonstrates we must constantly take new ideas forwards into trials but novel findings backwards to better understand what we already know. The clinician will be familiar with the language 'to translate from the bench to bedside and back again'. Understanding and employing bioinformatics to collect, integrate and present data,

ensuring that every step in this process is reproducible, is how we can avoid getting lost in translation.

References

- [1] An Overview of the Human Genome Project. Natl Hum Genome Res Inst NHGRI n.d. <https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/> (accessed May 26, 2018).
- [2] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. *Int J Inf Manag* 2015;35:137–44. doi:10.1016/j.ijinfomgt.2014.10.007.
- [3] Servant N, Roméjon J, Gestraud P, La Rosa P, Lucotte G, Lair S, et al. Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. *Front Genet* 2014;5. doi:10.3389/fgene.2014.00152.
- [4] Feldman MY. Reactions of Nucleic Acids and NucleoDroteins with Formaldehyde11Translated by A. L. Pumpiansky, Moscow. In: Davidson JN, Cohn WE, editors. *Prog. Nucleic Acid Res. Mol. Biol.*, vol. 13, Academic Press; 1973, p. 1–49. doi:10.1016/S0079-6603(08)60099-9.
- [5] Ludyga N, Grünwald B, Azimzadeh O, Englert S, Höfler H, Tapio S, et al. Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses. *Virchows Arch Int J Pathol* 2012;460:131–40. doi:10.1007/s00428-011-1184-9.
- [6] Robbe P, Popitsch N, Knight SJL, Antoniou P, Becq J, He M, et al. Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *Genet Med* 2018. doi:10.1038/gim.2017.241.
- [7] Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014;15:121–32. doi:10.1038/nrg3642.
- [8] Simon R, Roychowdhury S. Implementing personalized cancer genomics in clinical trials. *Nat Rev Drug Discov* 2013;12:358–69. doi:10.1038/nrd3979.
- [9] Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biol* 2011;12:125. doi:10.1186/gb-2011-12-8-125.
- [10] Tran B, Dancey JE, Kamel-Reid S, McPherson JD, Bedard PL, Brown AMK, et al. Cancer Genomics: Technology, Discovery, and Translation. *J Clin Oncol* 2012;30:647–60. doi:10.1200/JCO.2011.39.2316.
- [11] Yuan G-C, Cai L, Elowitz M, Enver T, Fan G, Guo G, et al. Challenges and emerging directions in single-cell analysis. *Genome Biol* 2017;18. doi:10.1186/s13059-017-1218-y.
- [12] Blainey PC, Quake SR. Dissecting genomic diversity, one cell at a time. *Nat Methods* 2014;11:19–21.
- [13] Koelzer VH, Gisler A, Hanhart JC, Griss J, Wagner SN, Willi N, et al. Digital image analysis improves precision of programmed death ligand 1 (PD-L1) scoring in cutaneous melanoma. *Histopathology* n.d.;0. doi:10.1111/his.13528.
- [14] Koelzer VH, Sokol L, Zahnd S, Christe L, Dawson H, Berger MD, et al. Digital analysis and epigenetic regulation of the signature of rejection in colorectal cancer. *Oncoimmunology* 2017;6. doi:10.1080/2162402X.2017.1288330.
- [15] de Bruijne M. Machine learning approaches in medical image analysis: From detection to diagnosis. *Med Image Anal* 2016;33:94–7. doi:10.1016/j.media.2016.06.032.
- [16] Tournéau CL, Delord J-P, Gonçalves A, Gavaille C, Dubot C, Isambert N, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol* 2015;16:1324–34. doi:10.1016/S1470-2045(15)00188-6.
- [17] Tsimberidou AM, Kurzrock R. Precision medicine: lessons learned from the SHIVA trial. *Lancet Oncol* 2015;16:e579–80. doi:10.1016/S1470-2045(15)00397-6.
- [18] McArthur GA, Chapman PB, Robert C, Larkin J, Haanen JB, Dummer R, et al. Safety and efficacy of vemurafenib in BRAFV600E and BRAFV600K mutation-positive melanoma (BRIM-3): extended follow-up of a phase 3, randomised, open-label study. *Lancet Oncol* 2014;15:323–32. doi:10.1016/S1470-2045(14)70012-9.

- [19] Prahallad A, Sun C, Huang S, Nicolantonio FD, Salazar R, Zecchin D, et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature* 2012;483:100–3. doi:10.1038/nature10868.
- [20] Masca NG, Hensor EM, Cornelius VR, Buffa FM, Marriott HM, Eales JM, et al. Science Forum: RIPOSTE: a framework for improving the design and analysis of laboratory-based research. *ELife* 2015;4:e05519. doi:10.7554/eLife.05519.
- [21] Baggerly KA, Coombes KR. What Information Should Be Required to Support Clinical “Omics” Publications? *Clin Chem* 2011;57:688–90. doi:10.1373/clinchem.2010.158618.
- [22] Read “Evolution of Translational Omics: Lessons Learned and the Path Forward” at NAP.edu. n.d. doi:10.17226/13297.
- [23] Suh KS, Sarojini S, Youssif M, Nalley K, Milinovicj N, Elloumi F, et al. Tissue Banking, Bioinformatics, and Electronic Medical Records: The Front-End Requirements for Personalized Medicine. *J Oncol* 2013;2013. doi:10.1155/2013/368751.