Fish genomes : a powerful tool to uncover new functional elements in vertebrates

Stupka, E.

Chapter 6: Discussion

Impact of next-generation sequencing on genome research

This thesis spans across several key genomics fields, reflecting the development of the discipline: from genome sequencing and assembly, to comparative genomics and transcriptomics. These fields have been impacted heavily by the emergence of next-generation sequencing, which has provided faster, more affordable tools to obtain genomes, transcriptomes, and "regulomes". In this discussion I aim to contextualize the results obtained during the thesis in the backdrop of these new technologies.

As noted by Lincoln Stein in his recent paper on cloud computing [1], while the cost of sequencing a base has fallen by half about every five months, the cost of storing each byte of data is dropping, but not as fast (every 14 months). This shifts completely the "data paradigm": while in the past obtaining data (e.g. the sequence of a gene) was a major achievement that would be closely guarded until publication, now obtaining data is easy, and the bottleneck becomes the bioinformatics analysis. Immediate, open, public release of data should thus be encouraged further, to enhance the data analysis potential and the biological conclusions that can be derived. This also has strong implications with regards to the type of biological questions that can be asked and the way in which they are asked. One obvious example is the genetics of outbred populations, unthinkable until recently, and now a reality [2].

The sequencing and assembly of entire genomes has been "commoditized" owing to next-generation sequencing. The effort involved in sequencing, assembling and annotating the Fugu genome in terms of money (approximately 10M

dollars), time (approximately 3 years) and people (approximately 20) involved would now require probably 2 orders of magnitude less effort (100K dollars, 3 months, 2 people). This reduction in costs/effort has brought the possibility of sequencing a genome to being a standard research project of a standard research lab. As an example, our lab is now actively involved in the genome sequencing of 4 species, and in the planning phases for approximately 25 more species, while a center such as the BGI in Shenzhen has recently set out to sequence 10,000 animal genomes.

The advances in data generation have pushed even further the key bottleneck towards the analysis of the data, i.e. the ability to use bioinformatics and biostatistics approaches to derive meaning from these large biological datasets. Moreover the data is often so rich that more than one person/team can utilize the same dataset and derive different, complementary, biological conclusions. One example is RNA-Seq, where the same dataset can be used to study several different biological aspects such as quantification of gene expression, discovery of novel genes, identification of alternative splicing. Each application requires different algorithmic approaches, dedicated bioinformatics effort and extensive validation. Thus the emphasis shifts away from being able to generate data to that of being able to analyze, obtain and validate novel biology reliably and extensively.

Searching for regulatory elements

The identification of regulatory elements genome-wide was, until very recently, confined to methods employing comparative genomics, such as the one looking we employed to identify shuffled conserved elements present in fish genomes,

presented in chapter 2. In this field, yet again, next-generation sequencing has had a major impact. The possibility to obtain genome-wide mapping of immunoprecipitated chromatin using Chip-Seq [3] has enabled, especially in mammalian systems, to obtain quickly very comprehensive signature of the regulatory code of the genome, including several histone marks, POL2 occupancy, and the regions bound by key de-acetylases such as p300, CBP [4]. Specifically in relation to enhancers a comprehensive studies conducted by Len Pennacchio showed clearly that a p300 Chip-Seq based approach [5] recovered with much higher success rates true functional enhancers than the older-fashioned sequence conservation based approach [6,7].

While Chip-Seq based approaches are clearly very promising, they are not directly and easily applicable to a large variety of species, as demonstrated by the fact that in non-mammalian vertebrate species, e.g. fish, so far there is no published extensive catalogue of enhancers. This is mainly due to the fact that the technologies need to be adapted to each specific species: the identification of antibodies that work effectively is often not trivial (easier for histone marks which are well conserved over longer evolutionary distances, but less straight-forward for DNA-binding proteins), the immunoprecipitation protocol needs to be adapted and optimized, and, last but not least, these techniques rely on large numbers of cells, which are often not available (and cell lines are often also not available). Finally, comparative genomics provides a large, unbiased, picture of regions of the genome that are under evolutionary constraint, regardless of their function. In order to identify all these regions via Chip-Seq approaches one would have to combine a vast number of Chip-Seq protocols, and might still miss

some with novel functions. Thus, for the time being, comparative genomics will still provide useful information on functional elements of the genome, which is complementary to Chip-Seq approaches.

Transcriptomics

In our study of MBT transition we had to resort to the technological platforms which were widely available at the time, i.e. microarrays. Microarrays have proven a fairly reliable measure of gene expression (for genes which are expressed at reasonable levels) in organisms such as the mouse and human genome. These organisms have benefited early on from a fairly complete genome sequence and assembly, very extensive biological sequence collections (ESTs, cDNAs, CAGE data, etc) and therefore gene annotation is very mature. This in turn has allowed microarray manufacturers to produce reliable oligonucleotide probes. Furthermore the very large market, usage and competition has forced continuous improvements of microarray platforms. The same cannot be said for other species. While many model organisms are not catered for at all by mainstream microarray manufacturers, for others, like *Danio rerio*, microarrays are available but far from ideal due to the poor (until recently) genome sequence and assembly and poor (until recently) gene models available. This is why in our analysis of MBT transition in zebrafish we could work only on approximately 10,000 genes, from which less than 2,000 were then usable for the final analysis.

RNA-Seq, on the other hand, provides a species-independent, unbiased, quantitative assessment of the transcriptome, which allows any lab, working on any species, to sequence the cDNA obtained from any RNA sample of interest.

Besides freeing the researcher from the need of a supported dedicated platform for the species of interest, it also captures a wider dynamic range of transcription, from very poorly expressed transcripts, to very highly expressed transcripts, without the limits imposed by the optical read-out of microarrays [8]. Moreover, RNA-Seq can be used effectively to study not only quantification of transcripts, but also alternative splicing [9] and novel gene prediction [10].

RNA-Seq on the other hand, as for many next-generation sequencing techniques, provides novel and difficult challenges from the bioinformatics analysis point of view. Mapping of reads to the genome is more complex due to the presence of spliced reads, which map across distant regions in the genome. Several algorithms have been developed in recent times to account for this aspect, such as, for example, TopHat [11] and SplitSeek [12], but these only aid in improving the quality of the mapping, without providing a complete solution for gene prediction or alternative splicing prediction. Newer algorithms such as Cufflinks [10] and many under development as part of the RGASP competition, such as mGene (developed by the group of Gunnar Rätsch at the Friedrich Miescher Institute) provide a much more sophisticated usage of RNA-Seq data and genome sequence to model accurately splice junctions, gene models and alternative splicing, for both coding and non-coding genes.

Genome Assembly

In genome assembly probably more than in any other genomics field the impact of next-generation sequencing has been radical. The commoditization of sequencing, coupled with the improvement of algorithmic tools and the commoditization of servers with large memory and CPU power has enabled the

average laboratory to undertake independently a whole genome sequencing, de novo assembly and annotation project, which until recently was confined to large sequencing centres. As shown in the last chapter of the thesis, the field is shifting rapidly, and while work was being conducted on the chapter new tools were being developed which assisted us in the de novo assembly of the carp genome. Although the work still needs to be complemented by further sequencing to improve contiguity we have shown convincingly that we were able to produce an assembly which is likely to contain a significantly large portion of the carp transcriptome (probably more than 90%) as assessed on the basis of both known carp DNA sequences as well as our own RNA-Seq dataset.

Similarly annotation of a genome was a heavy undertaking which involved comparative genomics as well as very expensive Sanger-sequencing based EST sequqencing projects. It can now be completed in a few weeks with a few Illumina lanes of RNA-Seq material, providing a good baseline for a preliminary annotation. In both the RNA-Seq and the genome assembly approach it is clear that the length of the sequences is still a limiting factor. Obtaining truly complete gene models from RNA-Seq requires very high depth. This, in turn, requires to obtain RNA from a range of tissues, or to utilize normalization protocols, since usually highly expressed genes will be well assembled, while genes with lower expression will have lower coverage and thus will not be assembled well. Similarly, while the genome assembly is satisfactory for preliminary identification of genes, mapping to other genomes, etc. it does not provide good multi-genic contiguity and is thus greatly limited in terms of more in-depth analysis. To achieve this either very high depth is required or some

complementary data, e.g. BAC end Sanger reads. As the cost of next-generation sequencing keeps dropping and sequence length increases, the cost/benefit ratio of using complementary Sanger based datasets will change. As shown with the publication of the Panda Genome [13], a complete de novo assembly with Scaffold N50 of over 1GB from Illumina sequencing only is now possible, as long as one can afford very high depth sequencing (in their case over 100X of the genome).

References

22. Stein LD The case for cloud computing in genome informatics. *Genome Biology* 2010; 1(5):207. Epub 2010 May 5
23. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 2008; 3(10): e3376. doi:10.1371/journal.pone.0003376
24. Valouev, A et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods 2008;* 5:829–834
25. Barski A, Cuddapah S, Cui K, Roh T, Schones DE, Wang Z, Wei G, Chepelev I Zhao K. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 2007; 129: 823–837
26. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA ChIP-seq accurately predicts tissue-specific activity of enhancers *Nature* 2009; 457:854-859
27. Pennacchio, L. A. et al. In vivo enhancer analysis of human conserved non-coding sequences. 2006 *Nature*; 444:499–502
28. Visel, A. et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. 2008 *Nature Genet*ics; 40:158–160
29. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B Mapping and quantifying mammalian transcriptomes by RNA-Seq 2008 *Nature Methods*; 5(7):621-628
30. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB Alternative isoform regulation in human tissue transcriptomes 2008 *Nature*; 456:470-476
31. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation 2010 *Nature Biotechnology*; 28(5):511-515
32. Trapnell C, Pachter L, Salzberg SL TopHat: discovering splice junctions with RNA-Seq 2009 *Bioinformatics*; 25(9):1105-1111
33. Ameur A, Wetterbom A, Feuk L, Gyllensten U Global and unbiased detection of splice junctions from RNA-seq data 2010 *Genome Biology*

11:R34
34. Ruiqiang Li et al. The sequence and de novo assembly of the giant panda genome 2009 *Nature*; 463:311-317