Massive parallel sequencing in animal genetics: wherefroms and wheretos

M. Pérez-Enciso*†‡ and L. Ferretti*†‡

*Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, 08193, Bellaterra, Spain. †Institut de Recerca i Estudis Avançats de Catalunya, ICREA, Barcelona, Spain. ‡Centre for Research in Agricultural Genomics, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

Summary

Next generation sequencing (NGS) has revolutionized genomics research, making it difficult to overstate its impact on studies of Biology. NGS will immediately allow researchers working in non-mainstream species to obtain complete genomes together with a comprehensive catalogue of variants. In addition, RNA-seq will be a decisive way to annotate genes that cannot be predicted purely by computational or comparative approaches. Future applications include whole genome sequence association studies, as opposed to classical SNP-based association, and implementing this new source of information into breeding programmes. For these purposes, one of the main advantages of sequencing vs. genotyping is the possibility of identifying copy number variants. Currently, experimental design is a topic of utmost interest, and here we discuss some of the options available, including pools and reduced representation libraries. Although bioinformatics is still an important bottleneck, this limitation is only transient and should not deter animal geneticists from embracing these technologies.

Keywords Association mapping, genome selection, next generation sequencing, pooling.

Introduction

The impact of massive parallel sequencing or ‘next’ generation sequencing (NGS) in Biology, and hence in Animal Genetics, is difficult to overstate. It is a revolution comparable with the one that followed Sanger sequencing forty years ago. It is not just a dramatic increase in sequencing speed: it means a change in paradigm that obliges researchers, institutions and funding agencies alike. It is also a tremendous and passionate technological race worth millions. Although it is uncertain which, if any, of the extant technologies will prevail in the future, and whether they may be replaced by new technologies, one thing is certain: next generation sequencing is this generation sequencing.

Current NGS technologies provide a throughput which is at least 100 times that of classical Sanger sequencing (Mardis 2008; Ansorge 2009; Metzker 2009), and the technologies are quickly improving (Drmanac et al. 2010). Compared with standard sequencing, these technologies do not need cloning, i.e. they are basically shotgun sequencing and result in shorter sequences, currently from 50 to 400 bp. Further, it is promised that new, third generation single molecule sequencing will soon be available and should deliver complete genome sequence that is in time and pricing affordable for every research group. The goal of this short review, however, is not to provide a comprehensive technological review. Incidentally, this would be a review that would be outdated by the time of publication. Our aim was rather to list a few key areas in animal genetics that have been dramatically affected by NGS technologies and to identify future applications and some associated challenges.

A reasonable question to ask, nevertheless, is whether we really need so much more sequence. Our answer is yes, we do. There are important advantages in having full sequence rather than genotypes: removing SNP ascertainment bias, uncovering all extant variability, recovering the full unbiased demographic history of populations or setting the theoretical ground for a unified framework that combines coalescence and association mapping. This is to name but a few advantages. Additional applications involve RNA-seq, to quantify precisely the transcriptome and study
allele-specific expression (Cloonan & Grimmond 2008; Bloom et al. 2009), or ChIP-Seq, to characterize the epigenome genomewide (Pepke et al. 2009). Exciting new approaches like metagenomics will soon allow us to study the genetics of ecosystems (Blow 2008). Thus far, the genetic composition of a population has been studied in isolation, while it is clear that animals co-evolve with many other living organisms, each with their own genomes. An example of particular interest in animal breeding occurs in ruminants: the physiological performance of dairy cattle depends critically on rumen microflora.

In summary, having more sequence will be beneficial to animal genetics and in particular to our understanding of the genetic basis of complex traits. Nevertheless, it is also true that many studies would trade less sequence for a more targeted approach. For some population genomics studies it is far more relevant to have, say, 10 Mb from 100 individuals than a 1 Gb from a single individual. Although several commercial approaches are already available for sequence capture, this increases costs significantly and does not provide a 100% recovery rate of the targets. This is one of the areas where we should witness rapid improvements soon, provided that sequencing a whole genome remains more expensive than targeted resequencing and this may not last too long, at least for a shallow coverage.

**Genomes and variants for all**

The time when comprehensive genomic data were available only in a very few species is over. NGS has democratized genomic resources and will provide fast and complete data à la carte in any species, allowing catching up with model or human species. Importantly, NGS should help not only by allowing cheap de novo sequencing, but also by improving gene annotation via RNA-seq (Cloonan & Grimmond 2008; Wilhelm et al. 2008), one of the most challenging tasks in any new genome. This advantage is particularly relevant in aquaculture, where dozens of marine and freshwater species are of interest, when compared with only a few major livestock species. Aquaculture species in turn provide wonderful examples to study the footprint of domestication in the genome, because domestication is happening right now. Although this is not a general rule, many fish species have relatively small genomes, like catfish (1 Gb) or turbot (<800 Mb), although others like salmon or trout have much larger genomes and/or are polyploids. Other species, e.g. medaka (800 Mb) have been used as model species for quite a time; several highly divergent medaka lines exist. Complete genome resequencing of these lines will provide extremely valuable knowledge about the genetic architecture of complex traits.

Thus far, large scale polymorphism data have been obtained from genotyping SNPs, typically in solid support or microarrays. Usually, these SNPs are chosen on a set of constraints, typically the minimum allele frequency (MAF) is required to be above a certain minimum (Ramos et al. 2009). For instance, a minimum MAF may be required in more than one population simultaneously, and, finally, SNPs are chosen also because of their technical suitability to be genotyped with the platform of choice. More often than not, SNPs will have been chosen according to different criteria which are not necessarily known. This poses important problems, of which animal breeders are not necessarily aware (Nielsen et al. 2004; Clark et al. 2005). Consider Fig. 1, which shows a hypothetical but plausible phylogeny: assume a collection of SNPs has been ascertained by resequencing a subset of animals, the discovery panel represented by the open circles. The SNPs found are those mutations represented by the white and grey arrows; and those chosen are the grey arrows after setting a restriction in MAF. Here, we discard singletons, i.e. alleles that are found only once in the discovery panel. The chosen SNP set is then utilized to genotype two additional populations, panel 1 and panel 2, the grey and black circles in the figure respectively.

There are at least two important consequences of using a biased subset of SNPs. The first one is a biased estimate of DNA diversity. If the genotyped panel is phylogenetically close to the discovery panel, the estimate will be biased upwards, because most of the SNPs will also be segregating at intermediate frequencies (panel 1), but, if the populations are distant, diversity measures will be severely underestimated, because the SNPs are undiscovered (panel 2). A complete range of intermediate scenarios is certainly possible. The second important result is that discovery of causal mutations might also be impaired, because the disequilibrium between causal mutations, especially if these are rare, and genotyped SNPs, will also be biased. In fact, we have recently shown that, for association mapping purposes, a random choice of SNPs equally spaced along the genome without regard for frequency is a more robust strategy than any filtering on MAF (Ledur et al. 2009). Logically, full
sequencing allows complete discovery of polymorphisms and therefore removes the issue of ascertainment bias. An additional advantage of complete resequencing is that not only are SNPs retrieved, but also structural variants like copy number variants (CNVs) are also identified. Nonetheless, this poses more algorithmic problems than usual (Medvedev et al. 2009). A potential problem with NGS data, especially in pools and/or with low coverage, is that telling singletons from sequencing errors can be difficult.

A further interest of NGS is that a single technology may fulfill (almost) all needs, making life simpler and slowly replacing more noisy approaches like microarrays. Transcriptome resequencing (RNA-seq) is today a much more accurate measurement of the level of gene expression than that provided by microarrays, and it delivers near-complete information on splicing variants. In the future, it should also be a cheaper option. However, the identification of splicing variants poses important computational problems (Pepke et al. 2009). Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) allows characterizing the modification status of DNA, e.g. methylation, genome-wide. Therefore, NGS will provide a near-complete picture not only of variants but also of transcriptomes and epigenomes. The potential impact on our understanding of how animal genomes function is dramatic.

Bioinformatics as a transient bottleneck

With NGS technologies, we have traded quality for quantity and a big price reduction. Currently, sequencing at low coverage for a typical mammalian genome is economically affordable within the scale of a medium-sized grant. Many research groups, nevertheless, lack the bioinformatic skills and computer resources needed to fully benefit from these data. Bioinformatic and analytical tools have therefore been recognized as the main bottleneck of current and forthcoming technologies (Pop & Salzberg 2008; McPherson 2009). Therefore, the reduction in costs can be misleading: computer and bioinformatic analyses will be much more expensive in the future than they are now. They will require either subcontracting computation power, e.g. cloud computing, or host and maintain dedicated servers, ideally hiring specialized personnel.

The bioinformatic challenges differ depending on the application: de novo sequencing, re-sequencing and RNA-seq. From a computational point of view the de novo assembly is by far the most demanding stage. The interested reader can follow some recent reviews on this hot topic (Flicek & Birney 2009; Imelfort & Edwards 2009). De novo assembly can be performed confidently with long reads (i.e. 454 technology), although new algorithms allow users now to carry out assembly with short reads, i.e. Genome Analyzer (Illumina Inc., San Diego, CA, USA) or Solid (Applied Biosystems, Foster city, CA, USA) (Li et al. 2009). In this case, paired-end sequencing is essential, as it is for discovering structural variants. As an application of the short read de novo approach, the complete panda genome has recently been assembled using exclusively Genome Analyzer at a 56× coverage (Li et al. 2010). The CPU and memory requirements are much higher with shorter reads and much higher coverage is required than with Sanger or 454 reads. For the panda genome, for instance, 32 CPUs and 512 GB of RAM were employed.

In a re-sequencing project, the goal is normally to identify variants, and a reference genome is assumed to exist to carry out the alignment. Therefore, from a computational point of view, it is much simpler than assembly, as it is equivalent to an automatized BLAST procedure. There are two important caveats though: the massive amount of reads produced and the specific sequencing errors produced by these technologies (Li et al. 2008; Flicek & Birney 2009). A typical re-sequencing pipeline consists of three main steps: sequence filtering, alignment and polymorphism detection. The final list of variants called can vary dramatically depending on the choice of parameters at every step; and the user is cautioned to judiciously fine-tune the different software options. Although standard filtering is applied by the sequencer software, additional filtering can be performed by the researcher to remove low quality or very short reads, highly repetitive reads or reads that do not match the genome. This will make the whole process more reliable. As for alignment, the most demanding step, a critical parameter is the number of mismatches allowed. By default, several popular software packages (Li et al. 2008, 2009) allow for two mismatches in short reads, but this restriction can be a bit too stringent with short reads that are lengthening as technology improves (reads are soon to be 100 bp in the Genome Analyzer) and in highly polymorphic regions. Finally, polymorphism detection is the primary goal for most animal geneticists. Although initially much effort has been devoted to SNP detection, a currently very active area of research is how to detect structural variants (Medvedev et al. 2009). However, despite recent advances in software, many challenges remain, such as detection of small indels (~<20 bp). This is a topic of increasing relevance.

RNA-Seq can be used, as can microarrays, to quantify number of transcripts (Mortazavi et al. 2008), but the fundamental advantage over microarrays is the possibility with RNA-Seq of discovering new splicing variants. This is, in turn, the main bioinformatic challenge, i.e. alignment when splice junctions are not known. If that is the case, then having paired-end reads is strongly recommended. Although there is comparatively less specific software for RNA-seq than for other applications (Pepke et al. 2009), this is rapidly changing, because RNA-seq should soon replace microarray technology and because the software can be used to discover chromosomal rearrangements, such as those in tumoral tissues. Given the large number of ongoing cancer sequencing projects, we expect a correlated increase in improved algorithms.
Aside from speed, memory requirements or ease of use, the user should consider carefully all alignment options to suit their needs, e.g. does it allow for gaps? How many mismatches per cluster are allowed? Will it automatically remove multiple matching or highly repetitive reads? It is also important to remember that both software and option choices will have a fundamental impact on the variant discovery results. This is especially true at low coverage, which is quite common, especially in early works that apply NGS. Visualization tools (Huang & Marth 2008; Skinner et al. 2009) do help, although logically these cannot be used to scan throughout all putative variants. All in all, the inexperienced researcher should not be discouraged by bioinformatic issues, as these will be transient. NGS will soon be standard routine in all genetics laboratories as software becomes more user-friendly than it currently is.

**Experimental design**

Although NGS has dramatically slashed the price of sequencing, we are still far away from the enticing $1000 genome, a price that would not cover the basic bioinformatic analysis in any case. It is necessary then to come up with cheaper, but probably also useful, alternatives in population genomics to individual sequencing. Non-mutually exclusive alternatives include pools, reduced representation libraries (RRL) or shotgun low fold coverage in a larger number of individuals. An RRL is made by digesting genomic DNA from a pool of individuals with a restriction enzyme and selecting a basepair range which is subsequently sequenced. The bottom figure is the opposite situation; it shows the probability of missing an allele as a function of true frequency; although as expected the probability represents the probability of missing an allele as a function of coverage (cov). For low coverage, the sequenced chromosomes are almost always distinct; by contrast, for high coverage, there are almost always chromosomes at least duplicated; finally, it is quite unlikely that all 58 chromosomes are represented at least once, even at very high coverage. For instance, the first line (5× coverage) shows that it is highly likely (P = 0.80) that each of the five sequences is originated from distinct chromosomes. At higher coverage (40×), only 30 origins will be represented on average. The middle figure represents the probability of missing an allele as a function of true frequency: although as expected the probability decreases with coverage, true singletons will be very difficult to spot. The bottom figure is the opposite situation; it shows the probability of sequencing an allele more than once, which is an approximate linear function of coverage and frequency. All in all, these graphs show that pooling will be a cost effective strategy for detecting SNPs, but that they will be biased towards mid frequency alleles; in general, it will be difficult to tell singletons from low-frequency alleles. Moreover, sequencing errors are more difficult to detect, and as a result, the variance of the estimators will be higher than in the individual sequencing setup. These considerations should be taken into account when pools are used to infer levels of variability.

The inference of demographic and selective processes is a fundamental problem in Genetics, and NGS applications in pools are discussed in (Ferretti et al. 2009 [see also Hellmann et al. 2008; Lynch 2008; Jiang et al. 2009 for inferences in individual sequencing]). Briefly, haplotype information is lost in the pooling process, at least for short sequences; therefore, inference has to be based on the SNP frequencies — site frequency spectrum or SFS — as a surrogate for levels of variability. The allele frequencies obtained from pooling correlate well with the frequencies in the whole population, at least for large pools and intermediate/high coverage. As illustrated in Fig. 3, ensuring that SNP-based inference from pooled data can be almost as powerful as from individual
towards intermediate-frequency alleles incurred by the SNP ascertainment process. Moreover, the resolution that can be achieved with RRL is higher than the current resolution of SNP chips for many species. Therefore, by estimating nucleotide variability in pools of extreme breeds, we can infer regions that putatively have been the target of selection or of domestication in a cost effective, genome-wide manner. Such an approach has been successfully applied in the pig (Amaral et al. 2009a). These kind of studies should rapidly replace current large scale SNP genotyping.

**Whole genome sequence association studies and sequence assisted selection**

Once locus discovery through linkage, i.e. pedigree-based analysis or QTL approach, proved to have insufficient accuracy, many studies turned to linkage disequilibrium instead. As a result, and also because of cheap genotyping costs, the number of published whole genome association studies (WGAS) has increased dramatically during the last years. Although partly successful, these studies suffer from the disadvantages of SNP ascertainment noted above, and are prone to a high false discovery rate. In addition, structural variants like CNVs are suspected to be associated with disease in many instances. Although it can be argued that SNPs in disequilibrium with structural variants could be used as surrogate variables, experimental data suggest that only a percentage of structural variants are tagged by SNPs (Redon et al. 2006). This is probably because their substitution rates and population dynamics are very different. To complicate matters further, the correct identification of structural variants is much more difficult than that of SNPs, particularly in the case of heterozygous individuals. Although the new generation of SNP chips are much denser and allow users to identify more reliably copy number variants than previously, the increased difficulty of reliably detecting structural variants relative to SNPs will persist no matter the advances in software.

A natural follow up to W Gas will be to carry out whole sequence association studies (WSAS). With current computers and NGS technology this may seem far away, yet we should be prepared to identify the main statistical challenges posed by large scale WSAS. A major one is coding the polymorphisms to establish a sound statistical model. This is illustrated in Fig. 4. In a typical WGAS, the phenotype is regressed onto each of the SNP genotypes in turn. More sophisticated statistical variants exist, but the key point is that each SNP is, in statistical terms, interchangeable. The model is the same throughout the genome. In a WSAS we could still use the same strategy, i.e. extracting the SNP list and proceeding as usual. However, we would be wasting much effort and discarding key advantages of using the sequence. With full sequence information, all kinds of variants are potentially identified, including indels, CNVs, and inversions, and SNP ascertainment issues are removed. As a result, modelling the relationship between sequence

---

**Figure 2** Pooling. We consider a pool of 29 individuals (58 chromosomes). Top: Number of distinct chromosomes sequenced as a function of coverage. Each line corresponds to a different number of origins, ranging from 5 to 40. Middle: Probability of not detecting an allele as a function of coverage and true number of allele copies in the pool. The number of copies ranges from $c = 1$ (uppermost line) to $c = 10$ (lowest line). Bottom: Probability of a given allele being sequenced more than once as a function of coverage and true number of allele copies $c$ in the pool. The number of copies ranges from $c = 10$ (uppermost line) to $c = 1$ (lowest line) in decreasing order.

---

Data. To infer SNP allele frequencies with a minimum reliability, it is recommended that a pool size of at least 10 individuals ($nc = 20$) is used and coverage of at least $20\times$ is obtained. Lower coverages and pool size result in rather uninformative SFS estimates (Fig. 3, top left). The availability of cheap sequencing in pools has immediate applications in livestock to infer demography structure and the footprint of selection, e.g. selective sweeps. Typical footprints of directional selection are regions of low variability, high linkage disequilibrium and an excess of low-frequency alleles. NGS shows a clear advantage over genotyping for the detection of low-frequency variants, because of the bias towards intermediate-frequency alleles incurred by the SNP

© 2010 The Authors, Animal Genetics © 2010 Stichting International Foundation for Animal Genetics, 41, 561–569
and phenotype is no longer as simple as with SNP-only information, meaning that not all genetic models are equivalent. This is illustrated in Fig. 4. Classical SNP data is represented in the top figure; in this case, a typical genetic model fitted to data is:

\[ y_i = \mathbf{X}_i \beta + \lambda q_j + e_i, \]

where \( y_i \) represents the \( i \)-th individual’s phenotype, vector \( \beta \) contains environmental effects, \( q_j \) the \( j \)-th additive SNP effect and \( e \), the residual. \( \lambda \) is an indicator variable that links genotype of individual with SNP effect. The same model (1) can be applied to every SNP, the only difference is the value of the indicator variable \( \lambda \), which varies according to genotype.

However, if sequence data is available, the scenario becomes much more complex (Fig. 4, bottom). First, we need to characterize variants: SNP, copy number variants (CNVs), indels and so on. CNVs in particular can have dramatic impacts on gene regulation and expression and are probably an important source of variability at the phenotypic level. Recent research shows that CNVs are very frequent and that up to 10% of the genome may be involved, and CNV maps in livestock are being rapidly developed (Fadista et al. 2008; Chen et al. 2009; Crooijmans et al. 2010; Du et al. 2010; Tang et al. 2010). Potentially, CNV models could be equivalent within polymorphism class, for instance, each copy number variant could be modelled similarly as in (1).

\[ y_i = \mathbf{X}_i \beta + \sum_h \sum_k \phi_{ikh} \alpha_k + e_i, \] (2)

where \( \phi_{ikh} \) is an indicator variable representing the CNV allele (e.g. number of copies) for the \( i \)-th individual, \( k \)-th CNV and \( h \)-th haplotype \( (h = 1, 2) \). Then separate models (1) and (2) could be fitted for SNP and CNVs. Aside from the fact of how to deal with all kinds of polymorphisms, at least two issues rapidly appear: i) how to consider SNPs within CNV stretches? A hierarchical model with nested SNP within CNV could be a solution, but this means that not all SNP models are equivalent; and ii) how do we consider potentially important interactions between and within variant classes? Competing models are not necessarily nested, making it impossible to use tests like likelihood ratio tests, so we need to resort to other criteria, e.g. Bayesian Information Criterion (BIC) and the like. Last but not least, all these specificities will require much faster algorithms than those currently available to explore a vast range of potential models over whole genomes. Therefore, WSAS will be feasible some time in the future in animals, but we do need to carefully prepare our statistical tools in order to not be disappointed and to extract the maximum amount of information.

A second topic of utmost interest in animal genetics is how to use molecular data to improve response to artificial
selection. There is certainly much excitement in the animal breeding community about genomic selection (GS), i.e. a procedure whereby animals are selected for using a multivariate regression-like procedure on thousands of markers (SNPs). Again, we should envisage a future where sequence rather than genotypes are the available information. One important consequence of sequence availability is that the classical distinction between identity by descent and identity by state will vanish. Nevertheless, it is doubtful that, even in a distant future, complete genome sequence is available for all individuals in a breeding programme. It is more likely that only offspring from elite sires are sequenced at low coverage. In a latter analysis, these data will be combined with imputation techniques such as those employed in the 1000 genome project as proposed by Howie et al. (2009).

In addition to bioinformatic challenges to automatize sequence data analysis, there remains much statistical and simulation work to be done. For instance, despite its relevance, we still do not know the impact of SNP ascertainment bias either on WGAS or on GS: extensive simulation research is badly needed to gauge the potential advantages of including sequence instead of genotypes under a range of plausible genetic architectures. Fast tools are also needed to carry out sequence association studies together with associated simulations to calibrate their power and false discovery rate.

Final thoughts

Every single corner of Genetics is being affected by these technologies. NGS has democratized genomic tools for labs that are independent or work in non-mainstream species. Therefore, NGS will foster a multipolar research where the influence of large sequencing centers will not be so prominent in setting the priorities of genomes to be delivered. Funding agencies should react accordingly. However, core centers will always be necessary for testing point technology, developing central bioinformatic applications and setting standards.

NGS technologies have accelerated and there is an ongoing and already increasing demand for bioinformatically oriented biologists. Programming skills are needed more than ever at a time when available software cannot fulfill all specific needs for every problem. PhD education should accompany these technological developments, and every PhD student in genomics should have a strong training in bioinformatics.

Although NGS applications to animal genetics are in their infancy, so far mainly to cheaply obtain thousands of SNPs, many more applications are being rapidly developed. From a technological point of view, the main constraints are targeted sequencing and multiple tagging rather than total sequence throughput. Nevertheless, a likely future is that sequence capture becomes more expensive than full resequencing. Bioinformatics is currently an issue to be considered, and is even limiting, but this transient bottleneck should not deter animal geneticists from applying these technologies in their research: exciting results await us just around the corner.

Acknowledgements

We would like to thank S. Ramos-Onsins, M. Groenen, H.J. Megens and A. Amaral for many discussions on this topic. We are grateful to reviewers’ help, which significantly improved the manuscript. Work funded by MICINN grant AGL2007-65563-C02-01/GAN (Spain).

References


© 2010 The Authors, Animal Genetics © 2010 Stichting International Foundation for Animal Genetics, 41, 561–569
