Next generation sequencing: Current status and prospects

David W. Burt
The Roslin Institute and R(D)SVS, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK
Dave.Burt@roslin.ed.ac.uk

During the 1990's genetic maps, mostly based on microsatellite DNA markers, were being developed in all the major livestock species, including poultry (Burt 2006). These maps were initially created to enable major QTL ("quantitative trait loci") for a wide range of traits to be mapped and exploited through marker-assisted-selection strategies. However, it soon became clear that a number of genomic resources were required before any significant progress could be made. This included the generation of genetic and physical maps, EST and gene expression arrays, etc (Burt 2006, 2007). The latter was the first large scale sequencing programme in the chicken, using Sanger or first generation sequencing technology, to sequence over 300,000 ESTs (Boardman et al 2002; Smith et al 2004). This was an exciting period when the human and other model genome sequences were being determined. The chicken genome was completed first (Hillier et al 2004), together with a set of three million SNPs segregating between broiler and layer lines (Wong et al 2004). Using the same methods the genome of the zebrafinch, a popular model in neurobiology, was completed later (Warren et al 2010) providing a comparison of two avian species. This was the start of the genome era in birds. Most of the resources, such as ESTs, arrays, SNPs etc were made available to the wider community through the ARK-genomics facility (www.ark-genomiocs.org). However these early sequencing projects required factory scale facilities, were very labour intensive and expensive projects (estimates are £1 million for the EST and £10 million for the genome sequencing projects). There needed to be a better way to sequence DNA, thankfully the human genome project provided the drive and next (2nd) generation sequencing ("NGS") technology came onto the scene (Table 1).

Table 1. Summary of sequencing technologies currently available.

NGS technology	Sequencing Principle	Read length in bases	% raw read accuracy	Reads per run	Gbases
1st generation					
Sanger	Dideoxy sequencing	~1,000	≥99.999	384	0.0004
2nd generation					
Roche/454	Pyrosequencing	350-450	≥99	8.00E+05	0.4
Illumina/Solexa	Reversible terminators	36–100	≥98–99	6.00E+09	600
ABI/SOLiD	Sequencing by ligation	35-60	≥99.99	1.00E+08	50–120
3rd generation					
Pac Bio	Single-molecule	1000-4500	≥80	4.80E+04	0.05
Helicos	Single-molecule	25–55	≥97	6.00E+08	21–35

Initially pyrosequencing (Roche/454) provided a system for the parallel sequencing of ~1 million DNA molecules with no cloning steps. It was cheaper, but reads were not so long as Sanger sequencing (Table 1). Currently, it produces the longest and most accurate sequence reads of any NGS method and is ideal for transcriptome assemblies. Other NGS methods are able to produce more data such as ABI/SOLiD and Illumina/Solexa, which are based on DNA

ligation and DNA synthesis, respectively. They are capable of sequencing billions of DNA molecules, generating short reads of less than 100 nucleotides, 100's of Gigabases of total sequence and lower error rates (Table 1). When combined with cheap disc storage of 100's terabytes, a terabyte RAM and 100-1000's of core processors, sequencing has become a tool, as well as a way to sequence genomes de novo (Ye et al 2011). These technology developments have been matched by a period of rapid software development for assembly of short reads into genomes or transcriptomes and the identification of genetic variants, such as SNPs and indels (Minou 2010).

Since the genomes of chicken and zebrafinch were completed, NGS has generated two more genomes each in less than a year, the turkey (Dalloul et al 2010) and recently the duck (Burt, unpublished), at a fraction of the cost. Currently the cost of a genome assembly is 25 USD (Ye et al 2011) and falling. Recent plans for 10K species (www.genome10k.org), includes at least 50 avian genomes (Jarvis, personal communication). The main focus of this collection is to solve the deep roots of the avian tree of life, often called the "wall of death". Analysis of available vertebrate genomes, including the four birds reveals patterns of gene expansions and contractions, as well rates of protein evolution which may be related to specific differences between birds and mammals, as well as lineage specific differences.

Resequencing of individuals can be used to catalogue the genetic variation between and within lines. As part of a large collaborative project involving many partners from academia (Roslin and Synbreed), animal breeding companies (Aviagen, Lohmann and Hy-line) and biotech (Affymetrix), we have sequenced 24 lines of poultry, as DNA pools of 10-15 animals. From this analysis we have defined 78M SNP segregating within these lines. The details of this work are described elsewhere in these conference proceedings (Boschiero et al 2001; Gheyas et al 2011). Such sequence data can also be used to define structural variants such as simple indels, to more complex inversions, translocations, CNVs etc. The frequency of SNPs can be used to define the relationship between lines. The distribution of SNPs along each chromosome can also be used to define chromosomal regions subject to selection, so called "signatures of selection". The SNPs will be made available from dbSNP and the Ensembl genome viewer in 2012. We have also used this large collection of SNPs to design a new high-density (HD) genotyping tool based on 600K SNPs (Boschiero et al 2001; Gheyas et al 2011) suitable for high resolution genetic mapping studies (GWAS) and genome selection (GS) in both broiler and layer lines.

The early EST and genome sequencing projects provided necessary information on the location of transcripts to enable the design of a number of gene expression arrays (e.g. Affymetrix). NGS can also be used to measure gene expression directly by sequencing mRNA abundance as a digital signal (RNA-Seq). Early studies used Roche/454 but more recently the throughput of Illumina/Solexa has been used in RNA-Seq. This method can be used to both discover then structure of genes, both coding and non-coding and also used to quantify expression from short e.g. miRNA and longer e.g. mRNAs. Entire transcriptomes can be assembled even without a genome reference sequence using de novo assemblers (e.g. ABySS, Birol et al 2009). However it is more convenient to map RNA tags onto a reference genome in which both coding and non-coding exons can be defined. Often more successfully than the traditional homology models, that often fails to predict 5, and 3' UTR's. The challenge will be to collect these data from as many studies and tissues as possible to fully predict the transcribed regions of the chicken genome; this is the plan for the next few years. Once we have defined the full set of exons, data from RNA-Seq experiments can then be fully analysed for example for differential gene expression (e.g. EDGE, Jeffrey et al 2006). In

addition these new predictions will form the basis of a new community driven exon array for rapid gene expression studies, still likely to outperform the RNA-Seq studies in cost and speed, with much less reliance on super computer time!

There are many other applications in which DNA fragments are isolated for example using binding proteins for methylated DNA (e.g. medips, Chavez et al 2010), antibodies for modified histones and transcription factor bound to DNA binding sites (e.g. ChIP-Seq, Liu et al 2010). Applications are only limited by our ingenuity and imagination of the researcher to isolate such sequences.

So what of the future? Currently there is much effort to develop yet another third generation sequencing technology based on single molecule sequencing, promising to be very fast, and providing even cheaper and longer sequences (e.g. PacBio and Helicos Table 1). The ultimate goal is to produce the 1000 dollar genome, which is not too far off. The ultimate for breeding will be routine sequencing of individual genomes; currently we can do this for a limited number of founder animals, and then integrate results with HD to low density SNP panels and imputation to predict the sequences of all animals (Howie et al., 2009). One day we will do this directly using NGS technology!

BIROL, I., S. D. JACKMAN, C. B. NIELSEN, J. Q. QIAN, R. VARHOL, G. STAZYK, R. D. MORIN, Y. ZHAO, M. HIRST, J. E. SCHEIN, D. E. HORSMAN, J. M. CONNORS, R. D. GASCOYNE, M. A. MARRA and S. J. JONES (2009) De novo transcriptome assembly with ABySS. Bioinformatics **25:** 2872-2877.

BOARDMAN, P.E., J. SANZ-EZQUERRO, I. M. OVERTON, D. W. BURT, E. BOSCH, W. T. FONG, C. TICKLE, W. R. BROWN, S. A. WILSON and S. J. HUBBARD (2002) A comprehensive collection of chicken cDNAs. Current Biology **12:**1965-1969.

BOSCHIERO C., A.A. GHEYAS, A. KRANIS, F.S. TURNER, K. WATSON, L. YU, R. TALBOT, S. SMITH, S.R. BROWN, P. KAISER, P.M. HOCKING, J.E. FULTON, T.M. STROM, G. HABERER, H.R. FRIES, S. WEIGEND, R. PREISINGER, S. QANBARI, H. SIMIANER, F. BREW, J.A.WOOLLIAMS and D.W. BURT (2011) Development of high density chicken genotyping array: I. detection of SNPs and criteria for SNP selection. Proceedings of the 7th European Symposium on Poultry Genetics.

BURT D. W. (2006) The chicken genome. Genome Dynamics **2**:123-137.

BURT D. W. (2007) Emergence of the chicken as a model organism: implications for agriculture and biology. Poultry Science **86:**1460-1471.

CHAVEZ, L., J. JOZEFCZUK, C. GRIMM, J. DIETRICH, B. TIMMERMANN, H. LEHRACH, R. HERWIG and J. ADJAYE (2010) Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. Genome Research **20:**1441-1450.

DALLOUL R.A., J. A. LONG, A. V. ZIMI, L. ASLAM, K. BEAL, L. A. BLOMBERG, P. BOUFFARD, D. W. BURT et al. (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. PLoS Biology **8**: e1000475.

GHEYAS A. A., C. BOSCHIERO, F.S. TURNER, A. KRANIS, K. WATSON, L. YU, R. TALBOT, S. SMITH, S.R. BROWN, P. KAISER, P.M. HOCKING, J.E. FULTON, T.M. STROM, G. HABERER, H.R. FRIES, S. WEIGEND, R. PREISINGER, S. QANBARI, H. SIMIANER, F. BREW, J.A.WOOLLIAMS, and D.W. BURT (2011) Development of high density chicken genotyping array: II. Validation of SNPs for final array design. Proceedings of the 7th European Symposium on Poultry Genetics.

HILLIER, L. W., W. MILLER, E. BIRNEY, W. WARREN, R. C. HARDISON et al. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**: 695-716.

HOWIE, B. N., P. DONNELLY and J. MARCHINI (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5: e1000529.

JEFFREY T., LEEK JT, E. MONSEN, A. R. DABNEY and J. D. STOREY (2006) EDGE: extraction and analysis of differential gene expression. Bioinformatics **22:** 507-508.

LIU, E. T., S. POTT and M. HUSS (2010) Q&A: ChIP-seq technologies and the study of gene regulation. BMC Biology **8**:56.

Minou N. (2010) Next-Generation Sequencing Techniques for Eukaryotic Microorganisms: Sequencing-Based Solutions to Biological Problems. Eukaryotic Cell **9**: 1300-1310.

SMITH, J., D. SPEED, A. S. LAW, E. J. GLASS and D. W. BURT (2004) In-silico identification of chicken immune-related genes. Immunogenetics **56**:122-133.

WARREN W. C., D. F. CLAYTON, H. ELLEGREN, A. P. ARNOLD, L. W. HILLIER et al. (2010) The genome of a songbird. Nature **464**:757-762.

WONG G. K. , B. LIU, J. WANG, Y. ZHANG, X. YANG, Z. ZHANG et al. (2004) Genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature **432:**717-722.

YE L., L. W. HILLIER, P MINX, N THANE, D. P. LOCKE et al. (2011) A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biology **12**:R31.