

# BBSRC REVIEW OF NEXT GENERATION SEQUENCING

**Final Version** 

#### **EXECUTIVE SUMMARY**

- 1. Knowledge of the sequence of deoxyribonucleic acid (DNA), the molecule that stores the genetic information of almost all known living organisms, has revolutionised biology and driven a massive acceleration in research and development. The most commonly used approach for DNA sequence determination has, until recently, been the chain termination methodology developed by Fredrick Sanger in the 1970's. The demand for genome sequence data using the Sanger methodology drove the formation of industrial-scale sequencing centres (e.g. Wellcome Trust Sanger Institute) and collaboration on a global scale. Scientifically, great strides forward were made in research areas where genome sequences were available but, due to the scale of the task and the cost only a limited number of species were covered and work on other organisms, many of significant strategic importance, was limited.
- 2. Since the late 1990's, researchers in both academia and industry have revisited the concept of DNA sequencing, looking to resolve weakness in the established methodology and maximise the efficiency of sequence production. These efforts have yielded a number of step-changing approaches that have been rapidly commercialised by companies including Illumina-Solexa, Roche and Applied Biosystems. Available since 2005, these new sequencing machines are delivering significantly cheaper and faster sequence generation into a market that demonstrates massive demand.
- 3. The relative affordability of next generation sequencing (NGS) machines is democratising high throughput sequence generation, potentially putting it in the labs of many researchers rather than just a few. As such, NGS presents great opportunities and challenges for biological research, facilities management and funding. It was recognition of these features that led BBSRC's Strategy Advisory Board (SAB) to recommend a strategic review of the area.
- 4. The aim of the Review is to examine NGS technology development and potential impact on biological research; establish a view of the current availability of equipment in the UK and associated data storage and management approaches; identify bottlenecks in NGS approaches including informatics tools and bioinformatics infrastructures; and provide advice to SAB on activities and investment that may be required to ensure appropriate adoption of NGS approaches and data collection, analysis and sharing. As a part of the review process, an expert working group was established, to provide technical and strategic advice under the chairmanship of Professor Ottoline Leyser FRS (University of York, Chair BBSRC Skills and Careers Strategy Panel, Member of BBSRC SAB).
- 5. The conclusions of the BBSRC Review of Next Generation Sequencing are:
  - NGS technologies are considered to offer several advantages (technological, cost and speed) over the Sanger methodology. Driven by huge market potential, improvements (2G) and new approaches (3G) are subject to intense commercial development. Collaboration between industry and academia on pre-commercial prototypes would be mutually beneficial.
  - NGS is having a significant impact across the biosciences. Current applications include *de novo* sequencing, re-sequencing, epigenetics and metagenomics. Over the next 5 years, NGS approaches are expected to become more widespread and well-established and BBSRC has a role to play in facilitating their adoption by supporting 'taster sessions' and pilot studies.

- Key areas of research and development that will benefit from NGS include food security (e.g. plant and animal breeding, disease diagnostics and management), industrial biotechnology (e.g. drug discovery, biopharmaceutical developments) and medicine (personalised medicine, disease diagnostics and management). The technology will have both economic and societal impact and the use of data arising from some potential applications (e.g. personalised medicine) raises significant ethical and societal issues.
- There is currently an appropriate balance between the supply of NGS equipment and the demands of the BBSRC user community. However, this is a dynamic situation and one that could quickly change; technology availability will need to keep pace with demand.
- There is a 'mixed economy' of NGS provision in the UK. The Research Councils and other funders have established centralised facilities, HEI's have set up local services and there is a growing number of commercial services. A significant proportion of NGS technology currently lies outside of the centralised facilities. Coordination is required to establish best practice, facilitate knowledge exchange and, importantly, to derive maximum value from the investments made.
- TGAC will boost UK capacity for sequencing plant, animal and non-medical microbial genomes. Clarification of the range of activities that TGAC will provide to the research community is required. A TGAC technical assessment process should be adopted, wherever relevant, across BBSRC funding mechanisms.
- The roll-out of NGS raises a number of funding issues: equipment cost and rapid improvement in machine specification makes it difficult to determine the 'right time to buy'; democratisation will continue and requests to fund machines outside TGAC are anticipated; and there is a need for greater transparency in the requests made to BBSRC to obtain funding.
- The NGS data production pipeline is currently at an early stage of development. There is evidence that experimental design may be influenced by cost, while production bioinformatics platforms are in a state of flux and there is a serious gap in data analysis and interpretation capabilities. Large-scale infrastructures appear to be managing the current level of data submissions and are planning for future expansion.
- Primary challenges for NGS chemical and molecular biology methodologies are centred on the effective parallelisation of sample preparation, and increasing sensitivities to allow reduced sample sizes. Addressing these challenges could increase the range of applications and lead to cost savings.
- There is a shortage of suitable software for sequence assembly of NGS data and in some areas (e.g. metagenomics) it is highly deficient.
- The massive expansion of sequence data being produced by NGS technologies is presenting challenges for commodity storage and data transfer. While storage is generally dealt with locally, transfer of data relies on the HEI network infrastructure.
- There is a major gap in NGS data analysis and data interpretation, which is exacerbated by a lack of capacity in computational genomics and bioinformatics.
- The ELIXIR project, if substantively funded, should go a long way to ensuring that bioscience researchers have access to appropriate bioinformatics infrastructure to enable the management and sharing of genome sequence datasets. However,

considerable problems will arise in providing appropriate infrastructure if ELIXIR investments do not substantively materialise.

6. The recommendations of the BBSRC Review of Next Generation Sequencing are:

**Recommendation 1:** BBSRC should encourage collaboration and partnership between leading sequencing facilities (e.g. TGAC) and NGS equipment manufactures to facilitate refinement of, and early-stage access to, the very latest technologies.

**Recommendation 2:** BBSRC should facilitate the widespread adoption of NGS by supporting 'taster sessions' and pilot projects that will enable roll-out of the technology across the key areas of application (e.g. re-sequencing, gene expression, epigenetics, metagenomics) within its remit.

**Recommendation 3:** To ensure that economic and societal impact is derived from the application of NGS approaches, BBSRC should promote knowledge exchange between academic researchers at the forefront of developing and utilising the new technologies, and stakeholders in industry and government.

**Recommendation 4:** BBSRC should monitor the availability and use of NGS technology, in order to maintain an awareness of the balance between supply and demand.

**Recommendation 5:** BBSRC (working with other partners where appropriate) should facilitate the coordination of UK NGS activities, to establish best practice, facilitate knowledge exchange and, importantly, to derive maximum value from the investments made.

**Recommendation 6:** BBSRC should: (a) work with TGAC to clarify the range of activities that it will provide to the bioscience user community, and; (b) Adopt a technical assessment process for TGAC wherever relevant, across BBSRC funding mechanisms.

**Recommendation 7:** BBSRC should ask applicants to justify (i.e. on the basis of cost and quality) the choice of sequencing facility as a part of the research grant application process.

**Recommendation 8:** BBSRC should encourage responsive mode proposals in NGS chemical and molecular biology methodologies.

**Recommendation 9:** BBSRC should encourage proposals to fill the gap in sequence assembly algorithms via the Tools and Resources Development Fund and the Bioinformatics and Biological Resources Fund.

**Recommendation 10:** BBSRC should ensure that the growing needs of the bioscience research community (including NGS data storage, interrogation and transfer issues) are considered in the multi-stakeholder discussions on e-infrastructures and IT networking.

**Recommendation 11:** BBSRC should, working with other partners as appropriate, develop a programme of activities to boost capacity and capability in bioinformatics and computational genomics, utilising the mechanisms identified in paragraph 86 and drawing on the expertise and experience at the centralised facilities (e.g. TGAC).

**Recommendation 12:** BBSRC should continue in its strong support for ELIXIR as demonstrated by proactive engagement in the preparatory phase and the recent financial commitment to upgrade the EBI's data infrastructure.

#### **BACKGROUND TO THE REVIEW**

- 7. Knowledge of the sequence of deoxyribonucleic acid (DNA), the molecule that stores the genetic information of almost all known living organisms, has revolutionised biology and driven a massive acceleration in research and development. The most commonly used approach for DNA sequence determination has, until recently, been the chain termination methodology. This was published by Fredrick Sanger (University of Cambridge) and colleagues in 1977 and led to Sanger's second Nobel Prize for Chemistry in 1980. The Sanger sequencing methodology has remained conceptually unchanged for over 30 years, but in the decades following publication improvements were made in reaction expediency and speed (e.g. dye termination methodologies, increased read lengths), process parallelisation (e.g. capillary electrophoresis) and automation. Automated sequencing machines first became commercially available in 1987 (US company, Applied Biosystems) and the current mature technology (e.g. ABI3730xl) is considered to offer high quality, high-throughput DNA sequence generation.
- 8. Using the Sanger methodology, the sequencing of large genomes (e.g. humans, other mammals and plants) presented a huge challenge in terms of scale and cost. This drove the formation of industrial-scale sequencing centres (e.g. Wellcome Trust Sanger Institute) and collaboration on a global scale. Sequenced genomes arising from this approach included model species (e.g. *Drosophila*, 2000; *Arabidopsis*, 2000) and species of medical and economic importance (e.g. Human, 2001; *Oryza sativa* (rice), 2002; *Bos taurus* (cow), 2009). Scientifically, great strides forward were made in research areas where genome sequences were available but, with only a limited number of species covered, work on other organisms was limited. Outside of the large sequencing centres, smaller facilities might sequence individual genes or carry out routine construct checking to support molecular studies but, on the whole, access to higher throughput sequencing technology was beyond the reach of most individual scientists and many research areas.
- 9. Hand-in-hand with industrial-scale DNA sequence production came the essential requirement for an infrastructure to provide the information to individual researchers around the world. Centralised facilities such as the European Bioinformatics Institute (UK) and the National Center for Biotechnology Information (US) worked in partnership with the DNA sequencing centres to assemble computationally the genomes, annotate the data, maintain the data resource, and make it available to researchers, together with complementary tools for analysis. These bioinformatics centres also accepted data from smaller sequencing enterprises, making it available to the community as a whole, and adding value to all sequence generation by placing it in a few, linked repositories.
- 10. Since the late 1990's, researchers in both academia and industry have revisited the concept of DNA sequencing, looking to resolve weakness in the established methodology and maximise the efficiency of sequence production. BBSRC-funded research at the University of Cambridge made a significant contribution to these efforts which, collectively, yielded a number of step-changing approaches that have been rapidly commercialised. The major companies in this new market include Illumina, Roche, Applied Biosystems, Helico and Complete Genomics. In general, these technologies offer faster and much cheaper (cost-per base / cost-per-read) sequencing than the existing Sanger methodology, but with a shorter read length. Driven by huge market potential, research continues apace to refine these commercial products and develop wholly new, alternative approaches. The alternative approaches are being developed by recently established companies, including Oxford Nanopore Technologies and Pacific Biosciences.

- 11. The relative affordability of next generation sequencing (NGS) machines is democratising high throughput sequence generation, putting it in the labs of many researchers rather than just a few. On the one hand this creates great opportunity, sequencing more genomes more quickly, opening up new lines of research and revitalising others, and potentially deploying sequencing in new technological contexts. On the other hand, it presents significant challenges, including: the effective coordination of fragmented sequence production activities and targets; curation and interpretation of a massively expanded body of sequence data, and; the potential of sequencing outputs to overwhelm existing genome information technology systems.
- 12. As the UK's principal funder of research into plants, animals and non-medical microbes, the Biotechnology and Biological Sciences Research Council (BBSRC) is responsible for scanning the science base to identify new developments that impact upon its mission (Annex 1). This activity is undertaken in partnership with the research community, through BBSRC's Strategy Advisory Board and Strategy Advisory Panels, and via wider community consultation. The outputs, usually in the form of a published report, inform BBSRC's strategy development activities and ultimately the Council's investment programme.
- 13. Strategy Advisory Board (SAB) identified NGS as a major new technology of relevance to research within BBSRC's remit and recommended that the Council undertake a strategic review of the area. In making the recommendation the Board expressed a particular interest in gathering evidence on the impact and availability of NGS and any emerging bottlenecks (e.g. bioinformatics) and gaps (e.g. skills and training). SAB also noted existing BBSRC plans to establish a large-scale sequencing centre (The Genome Analysis Centre, Norwich **Box 1**) and the proposed substantial upgrading of European life science data management efforts being taken forward through the European Strategy Forum on Research Infrastructures (ESFRI) Roadmap.

#### **Box 1: The Genome Analysis Centre**

The Genome Analysis Centre (TGAC) is a new UK resource for large-scale sequencing of plants, animals and microbes, and associated bioinformatics capabilities. TGAC was inaugurated in July 2009 under the Directorship of Dr Jane Rogers, previously of the Sanger Institute. The Centre represents a significant investment in bioscience research infrastructure led by BBSRC in a funding partnership with the East of England Development Agency and Norfolk Local Authorities (Norfolk County Council, South Norfolk District Council, Norwich City Council and the Greater Norwich Development Partnership.

The Centre complements the <u>Sanger Institute</u> and the research oriented sequencing hubs funded by MRC and NERC, and will work in close partnership with the <u>European Bioinformatics Institute</u>. The scientific remit of TGAC primarily covers that of BBSRC although some work of a more biomedical or environmental nature may be undertaken in partnership with industry and/or other research funding bodies.

TGAC is currently undergoing a phase of scale-up activities; as this phase matures the centre is expected to provide a range of outputs for the community, drawing upon a mixed economy of high throughput sequencing and several leading NGS platforms. These outputs are expected to range from large whole-genome sequencing projects to smaller scale applications of new technologies, and to include a substantial programme of methodological and analytical development. There is also expected to be a strong emphasis on research targets pursued in industrial partnerships.

#### **REVIEW PROCESS**

14. BBSRC's mission is to support research, training, knowledge transfer and public engagement in the biological sciences. This mission sets the overall framework for any strategic review. The terms of reference of the Review of Next Generation Sequencing are presented in **Box 2**.

#### Box 2: Terms of Reference

Review developments in NGS and its potential impact on biological research (animals, plants, non-medical microbes);

Establish a view of the current availability of the NGS equipment in the UK and associated data storage and management arrangements;

Identify bottlenecks in NGS approaches (i.e. what issues need to be tackled to ensure that the full benefit of existing and future NGS approaches are realised), including chemistry methodologies, molecular biology methodologies, informatics tools and bioinformatics infrastructures;

Provide advice to Strategy Advisory Board on activities and investment that may be required to ensure appropriate adoption of NGS approaches and data collection, analysis and sharing.

- 15. An Expert Group, chaired by Professor Ottoline Leyser FRS (University of York; member of BBSRC Strategy Advisory Board and chair of the Bioscience Skills and Careers Strategy Advisory Panel), was established to provide technical and strategic advice. The membership (Annex 2) included expertise in DNA sequencing technology, bioinformatics and facilities management, together with food security (crop production and animal health) and industrial biotechnology. Representatives from BBSRC's Genome Analysis Centre, the Medical Research Council, the Natural Environment Research Council and the Wellcome Trust attended as observers.
- 16. The Expert Group met twice during the Review process; in June 2009, to identify informational needs, discuss the terms of reference and examine the proposed community consultation, and in September 2009 to consider the responses to the consultation and identify conclusions and recommendations.
- 17. The Expert Group was provided with a range of evidence including:
  - The forty-two responses to a community questionnaire, composed of thirty from UK universities, four from BBSRC Research Institutes, three from industry and five from other organisations. One response was from a large-scale sequencing centre and two from smaller-scale sequencing hubs. Forty-one responses were from the UK. The consultation questionnaire is at Annex 3 and the list of respondents' organisations is at Annex 4;
  - Details of current BBSRC grants relevant to DNA sequencing, including improvements to methodologies, contributions to large-scale sequencing programmes, bioinformatics and the application of NGS approaches (Annex 5);

- A paper from the European Bioinformatics Institute on the use of NGS and its impact on bioinformatics.
- 18. The review report is informed by the evidence above and the discussions of the Expert Group. It was drafted by BBSRC Officials (Lead, Dr A Collis) with input from the Expert Group Members and Chair. The Chair approved submission of the final draft of the report to SAB.

#### **EXCLUSIONS**

19. The BBSRC Review of NGS is potentially of interest to a wide range of stakeholders including funders (research councils and charities), government policymakers, scientific researchers and facility managers. It is positioned in such a way as to be accessible to all of these groups and is not an in-depth scientific review of developments in the field. For detailed technical reviews, readers are directed to the 2009/10 Nature Reviews Genetics series on NGS and its applications (see <a href="https://www.nature.com/reviews/genetics">www.nature.com/reviews/genetics</a>).

## TERM OF REFERENCE 1: REVIEW DEVELOPMENTS IN NGS AND ITS POTENTIAL IMPACT ON BIOLOGICAL RESEARCH (ANIMALS, PLANTS, NON-MEDICAL MICROBES).

20. This section describes the NGS technologies, how they overcome some of the technical limitations of the Sanger methodology and how they differ from each other. It then sets out the key areas of application of NGS technologies together with identification of research and development areas where the provision of faster and cheaper sequencing is expected to have a major impact.

#### Second and Third Generation Sequencing

- 21. The automated Sanger methodology is referred to as a 'first generation technology' and NGS technologies are essentially grouped into second generation (2G) and third generation (3G) approaches. Several 2G approaches are commercially available (e.g. Roche-454, Illumina-Solexa, Applied Biosystems-SOLiD) whilst all 3G platforms are near market, with the exception of the Helicos HeliScope, which is already available. The Pacific Bioscience 3G technology is due for release in 2010. The key distinctions between the Sanger methodology and these next generation approaches and between 2G and 3G technologies are described below.
- 22. Second generation platforms remove the *in vivo* bacterial cloning stage of the Sanger methodology by using either 'emulsion PCR' (Roche-454, Applied Biosystems-SOLiD) or 'bridge PCR' (Illumina) for target amplification. Both methods circumvent the problem of host related bias (but not PCR bias), as well as speeding up sample preparation. All 2G approaches use a 'cyclic array' process in which dense arrays of DNA features are derived by iterative cycles of enzymatic manipulation followed by image based data collection (Shendure & Ji, 2008¹). They also miniaturise the sequence derivation stage by using strepavidin beads (Roche-454), flow cells (Illumina) or glass surfaces (Applied Biosystems-SOLiD). All 2G technologies are engineered to be massively parallel in terms of operations and output. To date, biological applications have focused on *de novo* sequencing, re-sequencing and metagenomics.

<sup>1</sup> Shendure, J. & Ji, H. Next-generation DNA sequencing. Nature Biotechnology 26, 1135-1145 (2008)

- 23. The commercially available 2G approaches are subject to on-going refinements in hardware, software and supporting methodologies. Key areas subject to improvements include read-length, accuracy, parallelisation (throughput, densities, and sample parallelisation) and time-per-run. There is also a drive to port them onto bench-top formats to enable use in smaller centres or individual laboratories. Two of the new platforms (Illumina / Applied Biosystems-SOLiD) yield shorter read lengths than the Sanger methodology. This is problematic because some short reads often fail to map uniquely to the genome, making sequence assembly difficult, especially with repetitive sequences. New bioinformatics algorithms for *de novo* sequence construction are required and while some progress is being made in this area using BBSRC funding (see Annex 5: BB/G024650/1, Principal Investigator Bevan; BB/G024715/1, Principal Investigator Birney), consultation respondents considered that the lack of available algorithms was frustrating the roll out of 2G approaches. This bottleneck is explored further under term of reference 3.
- 24.3G technologies deploy a single molecule template approach and, as such, remove the copy error and bias associated with PCR amplification. They also avoid the cyclic array approach and thereby enable further massive parallelisation. Read-out methods include differential conductance across nanopores (Oxford Nanopore Technology) and single molecule real-time sequencing using Fluorescence Resonance Energy Transfer (Applied Biosystems) or zero-mode waveguide detectors (Pacific Biosciences Limited). technologies are generally at the 'near market' stage of commercial development and potentially offer greatly increased read-lengths, the opportunity to work with very small sample sizes (due to direct reading), faster generation of outputs and significant improvements in sensitivity, and accuracy. An ability to obtain very long sequence reads from a single template molecule was considered by many respondents to represent the most important foreseeable technological step change for genomic research. Major areas for the application of 3G methodologies are anticipated to be sequence-based genotyping methods, transcriptome sequencing and some specific re-sequencing approaches. One consultation respondent commented that the possibilities for these technologies were 'almost limitless'. In the medium term, the cost of 3G approaches is expected to remain high and throughput possibly low compared to 2G.
- 25. Few respondents commented on wholly new sequencing approaches that might emerge in the future and an analysis of the BBSRC portfolio of grants did not identify any relevant research projects. Some respondents did, however, highlight the potential application of transmission electron microscopy and further developments in optical mapping technologies as areas of future development, and some basic research is ongoing in the improvement of existing methods (**Annex 5** BB/E025013/1, Principal Investigator Mir).
- 26. Collectively, the consultation responses conveyed a strong sense of rapid, on-going technology development in both 2G and 3G platforms. Indeed, one respondent commented that the Illumina 2G platform had 'improved 10-fold in the last year'. This dynamic situation presents a unique opportunity for collaboration between industry and academia. Both parties stand to benefit, with the company receiving early stage information on technology deployment and utility in a range of different areas, and academics gaining new insights for their research programmes from having access to the very latest technology. It is recommended that the UK sequencing centres (e.g. TGAC) seek such partnerships with commercial suppliers in this rapidly changing area. It was noted that the BBSRC Tools and Resources Development Fund provides financial support to facilitate these types of interaction.

Conclusion 1: Next Generation Sequencing technologies are considered to offer several advantages (technological, cost and speed) over the Sanger methodology. Driven by huge market potential, improvements (2G) and new approaches (3G) are

subject to intense commercial development. Collaboration between industry and academia on pre-commercial prototypes would be mutually beneficial.

Recommendation 1: BBSRC should encourage collaboration and partnership between leading sequencing facilities (e.g. TGAC) and NGS equipment manufactures to facilitate refinement of, and early-stage access to, the very latest technologies.

#### Potential impact on biological research

- 27. There was a high level of consensus amongst consultation respondents concerning the current use and potential future impact of NGS technologies. Those areas most frequently cited were scale-up of existing approaches (*de novo* sequencing and resequencing), replacement of established technologies (gene expression and epigenetics analysis tools), and opening up new areas of study (metagenomics). All these areas cut across biological research domains and therefore the potential impact of NGS is huge. Technology deployment is at an early stage and respondents anticipated that over the next five years it would become progressively more widespread. However, this roll out is heavily reliant on the development of new bioinformatics tools (see terms of reference 3). The impact of NGS on each area is described in the following paragraphs.
- 28. De novo sequencing: NGS technologies are currently being used to sequence the small, simpler genomes of prokaryotes. There is an expansion into understudied species and generation of additional reference genomes. The use of NGS in the *de novo* sequencing of eukaryotes is currently expanding, with uptake in cDNA sequencing studies and more limited application in whole genome approaches (e.g. some application in sheep and wheat). Looking ahead, it is expected that NGS deployment will become common-place, yielding the genomes of many more prokaryotic and eukaryotic species of strategic importance. Respondents anticipated that both large-scale sequencing centres and small scale laboratories will be engaged in *de novo* sequencing.
- 29. Re-sequencing: Respondents considered that NGS will yield a massive explosion in resequencing at the genome level. Until recently, re-sequencing has been limited to studies on small numbers of genetic loci. NGS is currently being used to re-sequence bacterial isolates to provide a better understanding of genomic diversity and the evolution of pathogens, and to identify mutant alleles and structural variants in human, plant and animal genomes. 2G and 3G technologies are considered complementary in resequencing, with 2G approaches utilised for variant discovery by whole-genome resequencing and 3G approaches used for discovering large structural variants and haplotype blocks. Looking ahead, NGS re-sequencing will enable the identification of divergent regions of the genome (e.g. disease susceptibility), the genetic basis of phenotypic polymorphisms and loss/gain in genomic variation. Comparing the genomes of multiple individuals in a species will enable detailed characterisation of individual and/or sub-species variation linking quantitative variation in phenotype to genotype, and allowing an understanding of the molecular basis for adaptation. Finally, some respondents commented that NGS-fuelled re-sequencing would open up a new area of ecosystem genomics (e.g. livestock and associated bacterial populations).
- 30. Gene expression and epigenetics analysis tools: NGS is a popular method for high throughput analysis of gene expression and is in the process of replacing DNA microarrays as the method of choice. Advantages of using NGS (termed RNA-seq) are: (a) it can be used with any organism as it is not restricted to the probe set on the microarray; (b) it offers greater sensitivity and dynamic range; (c) it can assess the levels of non-coding / antisense transcripts, and; (d) it detects sequence and splice differences. NGS is also replacing arrays in chromatin immunoprecipitation studies (termed ChIP-

seq), a method used in epigenetics to study DNA-binding proteins, histone modifications and nucleosomes. The UK is considered a global leader in epigenetics and a recent BBSRC/MRC investment (£1.65M) at the Babraham Institute will deploy NGS to investigate the epigenetic basis of normal development and healthy ageing. NGS is also being used to profile small RNA populations (sRNA-seq) that influence the transcription and translation of protein coding RNAs, and to identify DNA methylation through bisulphite NGS.

- 31. Metagenomics: Many respondents identified metagenomics, the study of the collective genomes of microorganisms, as an avenue of research empowered by NGS technologies. Metagenomic approaches identify genomic signatures from mixed populations of microorganisms, many of which may be unculturable. These signatures can act as markers (e.g. in environmental monitoring) or be analysed for useful leads (e.g. bioprospecting). Numerous basic research areas are enabled by improving tools for metagenomic characterisation of microbial communities, and their dynamics, with particular interests centred on human (the Human Microbiome project²), gut, rumen and soil microbial ecosystems. Respondents also identified detection applications in forensic science, bioterrorism and food poisoning. Looking forward, metagenomics (and metatranscriptomics) was considered to be an area with great potential, but requiring bioinformatics developments to effectively manage and interrogate the data.
- 32. Respondents conveyed a sense of real excitement at the prospect of widespread adoption of NGS approaches. NGS is considered to be a massively empowering technology that will enable researchers to choose the most appropriate biological systems for their research rather than being restricted to a small range of 'model' organisms for many types of investigation. It will also allow research to address questions that were previously prohibitively expensive to tackle. The Expert Group recognised that facilitating the widespread adoption of NGS presented a number of challenges and were supportive of schemes to promote expansion of usage whilst managing the risk associated with adopting a new approach. The Group recognised the high value of 'taster sessions' and pilot studies, and considered the NERC managed mode NGS pilot project initiative to be exemplary. The TGAC Capacity and Capability Call (Annex 6) currently provides a mechanism for pilot-scale community access to NGS approaches, but an on-going need for access of this type can be expected after the Centre is fully established.

Conclusion 2: NGS is having a significant impact across the biosciences. Current applications include *de novo* sequencing, re-sequencing, epigenetics and metagenomics. Over the next 5 years, NGS approaches are expected to become more widespread and well established and BBSRC has a role to play in facilitating their adoption by supporting 'taster sessions' and pilot studies.

Recommendation 2: BBSRC should facilitate the widespread adoption of NGS by supporting 'taster sessions' and pilot projects that will enable roll-out of the technology across the key areas of application (e.g. re-sequencing, gene expression, epigenetics, metagenomics) within its remit.

33. Strong themes emerged from the consultation responses concerning areas of research and development that would benefit substantially from the application of NGS approaches. These included plant breeding, plant-microbe interactions, animal breeding, industrial biotechnology, understanding pathogenesis and personalised medicine/diagnostics. The paragraphs below draw upon the consultation responses and provide a view of how these areas will benefit from the application of NGS technologies.

\_

<sup>&</sup>lt;sup>2</sup> http://nihroadmap.nih.gov/hmp/

Whilst personalised medicine is not within the scientific remit of BBSRC, many consultation respondents commented on the impact of NGS on this area and it is therefore included for completeness and for information.

- 34. Plant breeding: The application of NGS in plant breeding will yield reference genomes for crop species including those with very large genomes (e.g. wheat). Re-sequencing of cultivated and wild variants will provide insights into domestication and adaptation to changing environments (e.g. climate change). Re-sequencing of multiple crop cultivars will facilitate the dissection of the genetic architecture of quantitative phenotypic traits; this will allow marker-assisted selection on a genome scale, an approached dubbed 'genomics assisted breeding'. It will also facilitate the incorporation of useful allelic variants from landraces and ecotypes into domestic varieties, thus helping to maintain the influx of novel genetic variation into elite breeding material. NGS-enabled predictive breeding is expected to advance efforts to adapt crops for changing environments. The large-scale genotyping of thousands of individual plants is already a possibility for commercial selective breeding programmes, and the computational intensity of such work can be expected to increase significantly. There is, therefore, scope for significant socio-economic impacts from NGS-powered advances in crop breeding.
- 35. Plant-microbe interactions: Metagenomic studies will make an important contribution to understanding how plants and microbes influence rhizosphere composition, how microbial community composition affects microbial pathogenesis (e.g. soil-borne pathogens) and how new and emerging crop diseases originate. The ability to sequence, rapidly, the genomes of microbial pathogens affecting agricultural production will help to identify outbreak strains more quickly and enable prompt intervention.
- 36. Animal breeding: Similar to plant breeding, the key opportunities lie in the identification of genomic variation and its exploitation in selective breeding for improved productivity and health traits. It could also speed up disease diagnosis and the development of vaccines. Successful deployment could lead to significant economic and societal impact.
- 37. Pathogens: NGS-scaled methods identification and comparison of genetic differences between related pathogens will help researchers to elucidate how a pathogen adapts to new host ranges or environments. These studies will inform risk assessments for new and emerging diseases. Pathogen genome and population genome information may be used to map, understand and control important human, animal and plant pathogens. Within five years, targeted intervention may become feasible for serious bacterial infections; it is anticipated that such techniques will first applied to clinical applications, but that they will also become available to veterinary medicine.
- 38. Industrial biotechnology: Metagenomics studies are expected to make a significant contribution to industrial biotechnology, through identification of metabolic pathways and novel secondary metabolites (e.g. antimicrobials) with applications in drug discovery and biopharmaceutical development.
- 39. Personalised medicine and diagnostics: NGS could lead to the routine sequencing of individuals (e.g. at birth, or at medical need) and thereby usher in an era of personalised medicine. Genetic mapping and dissection of complex diseases could improve the ability of medical practitioners to predict the propensity of individuals to develop specific diseases (i.e. 'risk genes') and lead to development of earlier, more accurate and more sensitive disease diagnostics. Knowledge of an individual patient's genome sequence could inform treatment regimes, identifying appropriate drug targets and susceptibilities and enabling higher efficacy drug selection. Of particular note, identifying the complex genetic determinants and corresponding efficacious drug targets for individual cancer cases has the potential to revolutionise diagnosis, intervention and prognosis. Clearly,

common-place access to information about individuals' genomes also raises considerable ethical and societal concerns regarding the use and mis-use of this information, including information ownership and privacy.

- 40. Respondents also taxonomy and systems biology as two broad research methodologies that would be considerably benefited by application of NGS approaches. In both cases, access to large, accurate NGS generated datasets would enable deeper, more comprehensive studies to be undertaken. In the case of systems biology, in particular, the scales of investigation enabled by NGS could be considered integral to the high-throughput investigative modes that will be necessary to make significant future progress.
- 41. The potential economic and social impact of NGS is considerable and, in line with BBSRC's mission to facilitate knowledge transfer and innovation, steps should be taken to ensure effective knowledge transfer between industry sectors, government policy makers and academic researchers at the forefront of rolling out the new technologies. This could be achieved through the organisation of appropriate networking events/workshops.

Conclusion 3: Key areas of research and development that will benefit from NGS include food security (e.g. plant and animal breeding, disease diagnostics and management), industrial biotechnology (e.g. drug discovery, biopharmaceutical developments) and medicine (personalised medicine, disease diagnostics and management). The technology will have both economic and societal impact and the use of data arising from some potential applications (e.g. personalised medicine) raises significant ethical and societal issues.

Recommendation 3: To ensure that economic and societal impact is derived from the application of NGS approaches, BBSRC should promote knowledge exchange between academic researchers at the forefront of developing and utilising the new technologies, and stakeholders in industry and government.

## TERM OF REFERENCE 2: ESTABLISH A VIEW OF THE CURRENT AVAILABILITY OF NGS EQUIPMENT IN THE UK AND ASSOCIATED DATA STORAGE AND MANAGEMENT ARRANGEMENTS.

- 42. This section describes how NGS technologies are being deployed and used in the UK. It covers technology availability, the mechanisms that exist for accessing the technology and the funding implications of different routes of access. It then goes on to examine aspects of the NGS data production pipeline covering experimental design, data management, standardisation and data sharing.
- 43. Some of the quantitative information on technology availability is drawn from the consultation and, as such, it carries a number of risks. Firstly, the consultation responses do not necessarily present a complete picture of UK NGS provision; secondly, as provision and use of NGS is currently expanding, the data is probably already out of date. It is therefore likely that the information presented in paragraphs 44-45 presents an underestimate of actual UK provision.

#### **CURRENT AVAILABILITY OF NGS EQUIPMENT IN THE UK**

44. Thirty-seven of the forty-three consultation responses indicated that they either used (termed 'NGS Users') NGS technologies in their research or provided NGS services to the research community. Of the respondents that did not use NGS, none stated that this

was due to a lack of access to, or availability of, the technology. The turnaround time for most sequencing jobs appeared to be in the region of four to six weeks, but for one facility it was three months. The Expert Group considered that this was broadly timely, but some Members were aware of sequencing requests that had taken significantly longer due to a facility suddenly facing overwhelming demand. Members hoped that the extra capacity delivered through TGAC will relieve the pressure on capacity.

- 45. The NGS Users obtained NGS derived sequence via a number of different routes and several respondents used more than one of these routes. Within the User Group, eighteen used UK sequencing hubs (eight Liverpool; eight Edinburgh; two WTSI), eighteen used facilities in their own institutions (including the sequencing hubs), ten used facilities outside of the hubs or their institutions (UK and overseas) and seven used commercial suppliers (UK and overseas). The User Group did not cite BBSRC's TGAC as the Centre had not initiated operations at the time of the consultation. Low usage of the Wellcome Trust Sanger Institute, one of the largest sequence facilities in the world, reflects the distinct difference between the Trust's mission and the research areas supported by BBSRC. Hubs funded by CRUK are also unlikely to be used by BBSRC funded researchers as they are restricted to researchers working on cancer.
- 46. Researchers who responded to the consultation held a variety of opinions on whether UK NGS provision equals, exceeds or trails behind that of other countries. Overall, the responses suggested that the UK appears to be around two years behind the USA, and somewhat ahead of most of Europe, with the possible exception of Germany and/or France. Several respondents noted that China was rapidly accumulating NGS capacity.
- 47. Based on the information provided in the consultation responses and with the recent launch of TGAC, the Expert Group concluded that there is currently an appropriate balance between the supply of NGS equipment and the demand from the BBSRC user community. However, the Group considered that this was a dynamic (perhaps unstable) situation and one that could quickly change. Whilst more machines will no doubt come on line, the number of NGS users is expected to increase significantly. The Expert Group recommended that BBSRC should monitor the availability of the technology.

Conclusion 4: There is currently an appropriate balance between the supply of NGS equipment and the demands of the BBSRC user community. However, this is a dynamic situation and one that could quickly change; technology availability will need to keep pace with demand.

Recommendation 4: BBSRC should monitor the availability and use of NGS technology, in order to maintain an awareness of the balance between supply and demand.

#### **NGS ACCESS - MANAGEMENT ARRANGEMENTS**

#### General Observations

48. Research Councils, charities and other funders have supported the deployment of NGS technologies in a similar way, although the scale of operations varies (**Table 1**). The model used in all cases is a centralised facility providing sequencing across a broad range of research areas and based within, or in close proximity to, a leading research university and / or research institute. Centralised facilities offer economies of scale, provide the most supportive environment for rolling out the very latest equipment, provide a hub for training activities and can lower the entry barriers for researchers adopting new technologies by providing access to expertise. Establishing a close association with life science researchers is seen as beneficial to a facility, helping it to remain relevant to the

changing needs of the research community and providing collaborative opportunities to develop new methods.

**Table 1**: Major UK sequencing facilities offering NGS to the research community (Autumn 2009).

Facility	Equipment	Major funders	Academic access
TGAC	2Illumina GAII 2 Roche 454 2 ABI SOLID 15 ABI 3730	BBSRC, EEDA, Norfolk Local Authorities	Costed on research grants
Edinburgh	2 Illumina GAII 1 Roche 454 2 ABI 3730	NERC, MRC <sup>1</sup>	Costed on research grants
Liverpool	2 Roche 454 2 ABI SOLiD	NERC, MRC	Costed on research grants
Sanger Institute	37 Illumina GAII 2 Roche 454 FLX 35 ABI 3730	Wellcome Trust	Institute identifies projects. Not open access.

<sup>&</sup>lt;sup>1</sup>MRC funded four NGS sequencing hubs in 2009/10. In addition to Edinburgh and Liverpool, hubs were established at Cambridge and Oxford. Total investment: £9.1M.

- 49. The centralised sequencing facilities (**Table 1**) are complemented by machines housed in individual HEI's and the overall picture of NGS deployment in the UK is of a 'mixed economy'. Indeed, if the Sanger Institute is excluded from consideration, it appears that individual HEIs have provided an equal or slightly greater level of NGS capacity than the centralised facilities. Respondents stated that institutional NGS facilities were in the process of being set up at the Universities of Birmingham, Bristol, Exeter, Manchester and Warwick. Sources of funding include the Funding Councils, the respective university and a Regional Development Agency. Where new facilities were being established, several respondents identified a preferred approach to deployment, with NGS being integrated into existing core facilities that provide services across a faculty or several departments. This appears to maximise machine usage and provides some dedicated expert technical support (analytical and informatic).
- 50. The democratisation of sequence production, reflected in the 'mixed economy' of current UK provision, presents a new challenge; of coordinating sequencing activities at differing scales and in different settings. The Expert Group considered that coordination was essential to establish best practice, facilitate knowledge exchange and, importantly, to derive maximum value from the investments made. This Review proposes that BBSRC facilitate networking of the centralised facilities, not only with each other, but also with the HEI funded services. Issues that could be covered by such networking include:
  - Experience of rolling out new equipment;
  - Procurement of machines and consumables:

- Establishing best practice throughout the sequence data generation pipeline including development of data standards;
- Provision of a forum for communication between a multitude of sequence producers and the centralised data sharing infrastructures (e.g. EBI); and
- Development of a collective view on the activities and specialisms of the sequencing services across the UK.

Conclusion 5: There is a 'mixed economy' of NGS provision in the UK. The Research Councils and other funders have established centralised facilities whilst HEI's have set up local services. A significant proportion of NGS technology currently lies outside of the centralised facilities. Coordination is required to establish best practice, facilitate knowledge exchange and, importantly, to derive maximum value from the investments made.

Recommendation 5: BBSRC (working with other partners where appropriate) should facilitate the coordination of UK NGS activities, to establish best practice, facilitate knowledge exchange and, importantly, to derive maximum value from the investments made.

#### TGAC

- 51. The consultation respondents and Expert Group broadly welcomed the creation of TGAC, a dedicated sequencing facility for plants, animals, and non-medical microbes. There was a consensus view that, over many years, a lack of capacity has limited the UK's contribution to sequencing of these organisms; the advent of NGS technologies, their broad availability and the launch of TGAC should enable considerable progress to be made towards remedying this longstanding issue. The Expert Group noted the equipment composition of each of the centres. TGAC's NGS provision is similar to, but pulling ahead of, the NERC/MRC hubs, and overall composition is heavily biased towards the ABI3730s.
- 52. Members queried the extent to which TGAC would operate in all of the application areas identified under term of reference 1. In particular, members questioned whether TGAC would undertake smaller scale and routine experimental applications of NGS (e.g. in gene expression analysis, ChIP-Seq studies, etc) or novel assay development for specialist research areas. The Expert Group considered that BBSRC should work with TGAC to clarify the scope of its activities and then communicate this message to the research community.
- 53. As a central facility providing a service to the research community, it is important to manage potential conflicts of interests between the individual researcher (i.e. user) and the facility (i.e. supplier). These may arise due to inappropriate technical requests, nature of the supplied input and output, and/or time for completion of the service. Members considered that a formal technical assessment, agreed between the user and supplier, is important in setting expectations and were pleased to note that TGAC had introduced such a process in the current Capacity and Capability Call (Annex 6). The Review recommends that a TGAC technical assessment process be introduced to BBSRC responsive mode and other relevant funding mechanisms.

Conclusion 6: TGAC will boost UK capacity for sequencing plant, animal and non-medical microbial genomes. Clarification of the range of activities that TGAC will provide to the research community is required. A TGAC technical assessment process should be adopted, wherever relevant, across BBSRC funding mechanisms.

Recommendation 6: BBSRC should: (a) work with TGAC to clarify the range of activities that it will provide to the bioscience user community, and; (b) Adopt a technical assessment process for TGAC wherever relevant, across BBSRC funding mechanisms.

#### NGS ACCESS - IMPLICATIONS FOR FUNDING

- 54. Significant machine costs (Roche 454, \$500K; Illumina, \$540K; Applied Biosystems-SOLiD \$595K) coupled with rapid technology upgrades presents a significant challenge to purchasers and funders of NGS equipment. In determining the right time to buy, purchasers need to balance current demand and availability against likely future availability of higher specification equipment. Buy too early and researchers can quickly fall behind those who adopt later technologies, buy too late and the research base may already have suffered from a loss of momentum. Technology obsolescence issues may arise quickly if the wholly new systems under development (e.g. 3G NGS) displace the current machines. The potential for 'quick-turnover' of expensive NGS equipment may present an extreme case for depreciation under full economic costing (fEC). An alternative approach to procure NGS, whilst managing the risks associated with purchase, is leasing. Some companies (e.g. Illumina) offer this service, but the Expert Group is aware of only a single example of its use in the UK (University of York).
- 55. The Expert Group considered that demand for local NGS services is likely to increase as the number of users expands and companies bring bench-top models to market. This could put pressure on the Research Councils to fund NGS machines in individual HEIs / institutes as well as the sequencing centres. None of the life science funders currently operates a dedicated mid-range / multi-user equipment funding scheme, with BBSRC curtailing the Research Equipment Initiative in 2007. At that time, the Council set out that the cost of replacing equipment should be covered by the directly allocated costs (i.e. via facilities charges, depreciation) on research grant applications, now submitted under fEC. However, the door was not entirely closed to mid-range equipment requests; advice was also issued stating that such requests could be made on responsive mode research grant applications, although few have been received. The Expert Group queried whether BBSRC's policy for mid-range equipment was adequate and welcomed the review of the Research Equipment Initiative (1999-2004) which is currently in progress. Members considered that BBSRC should consider applications for NGS technologies outside of TGAC, but only if the equipment will benefit a community of researchers.
- 56. A challenge for BBSRC, inherent in the 'mixed economy' of NGS provision, lies in the multiple routes through which the Council is currently asked to fund the cost of sequencing for the UK bioscience research community. The routes include funding of a national capability at TGAC (either as block grant to the centre or via charging on research grants); other services offering transparent charging (i.e. company contracts, MRC/NERC pay-as-you-go sequencing hubs costed on research grants); and the generally non-transparent, directly allocated budget line on research grants. multiplicity of routes and variable transparency in funding requests makes it difficult to determine which facilities offer the best value for money, provide the most viable and useful outputs and enable researchers to comply with data sharing commitments. In light of this situation, it is proposed that applications involving NGS should clearly and specifically justify their choice of NGS provider. The peer review committees can then consider whether the requests represent an appropriate use of BBSRC funds. It was noted that relevant information on TGAC, concerning costs, service and quality, will shortly be captured within the BBSRC electronic grant submission process, via incorporation of the TGAC technical assessment process.

Conclusion 7: The roll-out of NGS raises a number of funding issues: equipment cost and rapid improvement in machine specification makes it difficult to determine the 'right time to buy'; democratisation will continue and requests to fund machines and services outside TGAC are anticipated; and there is a need for greater transparency in the requests made to BBSRC to obtain funding.

Recommendation 7: BBSRC should ask applicants to justify (i.e. on the basis of cost and quality) the choice of sequencing facility as a part of the research grant application process.

#### THE NGS DATA PRODUCTION PIPELINE

57. This section examines key stages in the production of NGS-derived sequence data. It draws on the consultation responses to identify major themes emerging from the current approaches being used to roll-out NGS technologies. Term of reference 3 complements the information presented in this section by addressing the challenges and bottlenecks in adopting NGS sequencing.

#### Experimental design

- 58. The Expert Group recognised that it was difficult to provide general advice on what constituted robust experimental design as this was often dependent on the context within which NGS was being deployed. For example, one respondent working on DNA-protein binding stated, entirely appropriately, that 'fishing expeditions' were of 'no value', whilst for those working on metagenomics, controlled 'fishing expeditions' are the basic premise on which their research is based. One strong theme on experimental design did, however, emerge and that was around the need for appropriate biological replication and experimental approaches that facilitate robust statistical analysis.
- 59. Some respondents raised concerns that researchers were compromising technical and biological replication in a drive to keep costs down on grant applications. One commented that 'so far, very little published work has any real replication'. This situation was likened to the roll-out of microarray technology several years ago, when costs were minimised by using few arrays and pooling samples rather than running multiple samples separately. To emphasis the point, one respondent commented that 'lessons should be learned from thousands of badly designed microarray experiments.....such experiments are useless and only eat up resources'. The expert group agreed that robust statistical design should be an integral part of NGS experimental design and that it was essential for researchers to understand the appropriate requirements for biological and experimental replication, to give authority to experimental outputs. However, the Group also considered it important to recognise that the diversity of NGS applications makes it impossible to provide blanket recommendations; each experiment must be considered based on its specific aims. Again recalling the early microarray era, attempts to enforce unnecessary replication were often highly wasteful of resources.

#### Data storage and data management approaches

- 60. One of the challenges of NGS is the storage and management of the vast quantities of data generated (especially that arising from short-read platforms). The consultation gave some insights into the way researchers and sequencing services are responding to the challenges in the distributed, 'mixed economy', of sequence production that now exists.
- 61. When asked about the infrastructure arrangements in place for handling NGS-derived sequence data, respondents indicated that the production bioinformatics directly associated with sequencing (i.e. quality control, alignment, assembly) tend to be

undertaken in-house. Several respondents expressed concerns, however, that support was "poor", "make-shiff" or "underdeveloped". Others stated that they were currently setting up facilities, building laboratory Information management systems (LIMS), boosting commodity storage clusters and installing large servers or using local high performance computer facilities for data analysis. The different needs of the various sequencing platforms adds another level of complexity to the provision of IT support, as does the regular upgrades to hardware and software platforms issued by suppliers. These factors lead to in-house systems that are constantly in flux and require dedicated computer officer expertise to maintain and develop the IT infrastructure.

- 62. There was a considerable diversity of views expressed on the subject of whether standards or minimum information systems existed or were under development for reporting the outputs and outcomes of NGS experiments. Only three respondents cited relevant minimal-information reporting systems, including minimum information about a high throughput sequencing experiment (see www.mged.org/minseqe). The patchy awareness amongst respondents suggests under-engagement in the formulation and deployment of these systems. The network proposed under recommendation 5 could strength the UK's input into international standards activities as well as disseminate best practice across the fragmented community.
- 63. Respondents cited the NCBI and EBI databases (e.g. the short-read archive) as places where they deposit NGS data. These archives impose minimum standards on metadata associated with sequence data submission and several respondents considered that their involvement in the development of standards was essential. Some Members and respondents queried whether these facilities had the capacity to manage a massive influx of sequence data. Others considered that through initiatives such as the 1000 genome project<sup>3</sup>, data archiving capacity should be sufficient for the foreseeable future (2012).
- 64. Current approaches to sharing NGS data are largely in line with funders' policies on sharing research data and, in the majority of cases, data sharing takes place at the point of publication. This approach is different from that deployed in the large-scale genome sequencing projects (based on Bermuda and Fort Lauderdale guidance<sup>4</sup>) where release is required immediately after the sequence data had been derived. It is more in keeping with the practices in other research areas not involving sequencing.
- 65. A strong collective view arose from the consultation responses on the subject of data analysis and interpretation. This was recognised as a "big problem", a "limiting factor" and a "major challenge". With production informatics (paragraph 61), there was a sense that institutions were starting to tackle the challenges, but for data analysis and interpretation there appeared to be a major gap with the consultation responses providing little insight to the progress being made. The Expert Group considered the data analysis challenge to be immense and noted that it raises research, skills and infrastructure issues. Researchers will be unable to take full advantage of NGS in biological discovery and maximise the value of recent investments unless it is tackled. Tackling the challenge of data analysis is explored in depth under term of reference three.

Conclusion 8: The NGS data production pipeline is currently at an early stage of development. There is evidence that experimental design may be influenced by cost, while production bioinformatics platforms are in a state of flux and there is a serious gap in data analysis and interpretation capabilities. Large-scale infrastructures

\_

<sup>&</sup>lt;sup>3</sup> http://www.1000genomes.org/

<sup>4</sup> http://www.genome.gov/10506537

appear to be managing the current level of data submissions and are planning for future expansion.

TERM OF REFERENCE 3: IDENTIFY BOTTLENECKS IN NGS APPROACHES INCLUDING CHEMISTRY METHODOLOGIES, MOLECULAR BIOLOGY METHODOLOGIES, INFORMATICS TOOLS AND BIOINFORMATICS INFRASTRUCTURES

66. Building on the information under term of reference 2, this section sets out the technical challenges and bottlenecks that exist in current NGS approaches and the sequence data pipeline. It covers, where appropriate, research challenges, skills gaps and training needs, and the demands placed on large-scale infrastructures.

#### CHEMICAL AND MOLECULAR BIOLOGY METHODOLOGIES

- 67. Current challenges in the area of chemical and molecular biological methodologies centre predominantly on issues of biological sample preparation prior to sequencing, and increasing the speed, simplicity and reliability of these processes. Current protocols are technically challenging, and good quality and reproducibility is dependent on a high level of technical skill and experience. There are considerable potential cost savings associated with improving the robustness and efficiency of sample preparation protocols; labour and reagent costs in sample preparation work can be prohibitively high and place limitations on the application of NGS.
- 68. Numerous respondents said it was difficult to prepare enough samples in parallel to match the capacity of NGS machines. There are also current technical barriers to increasing sensitivity and reducing sample sizes reliably and thereby expanding the range of biological sources (e.g. facilitating single cell sampling).
- 69. For some approaches (e.g. RNA-Seq) it is clear that the development of optimised NGS protocols represents a particular ongoing challenge. More broadly, the process of protocol development, optimization and automation for the production and processing of samples is evidently a fast moving field. There is, therefore, a need for the efficient community dissemination of both best practice and innovations (specific examples include the development of molecular bar-coding and use of spiked controls).
- 70. There was a clear consensus from respondents that, whilst there are challenges for current and future chemical and molecular biology tool development, these do not represent the major barriers to the widespread adoption of NGS approaches. The most significant challenges lie further downstream in informatics derivation and interpretation of NGS data outputs. However, that said, the Expert Group considered that BBSRC should still encourage applications on technology development into responsive mode, continuing to build on the UK's strong reputation in sequencing chemistry and molecular biology.

Conclusion 9: Primary challenges for NGS chemical and molecular biology methodologies are centred on the effective parallelisation of sample preparation, and increasing sensitivities to allow reduced sample sizes. Addressing these challenges could increase the range of applications and lead to cost savings.

Recommendation 8: BBSRC should encourage responsive mode proposals in NGS chemical and molecular biology methodologies.

#### **INFORMATICS TOOLS AND INFRASTRUCTURES**

71. The advent of NGS has alleviated a major bottleneck for genomics research, namely the speed and cost (and allied accessibility) of the physical process of generating genome sequence data. However, this success creates new pressures in other parts of the pipeline from generating genome sequences to deriving new biological insight from them. Almost all of the consultation respondents recognised that NGS presents major challenges in the areas of bioinformatic derivation and interpretation, and in ensuring that appropriate bioinformatics infrastructures are in place to ensure that maximum value is derived from the sequence data generated.

#### Sequence assembly

- 72. Before it can be interpreted in a biologically meaningful way, sequence reads must be assembled to generate genome-scale sequence information. The complexity of this task is increased with 2G NGS approaches, which typically produce much shorter individual read lengths than Sanger sequencing methods, thus increasing the incidence of sequence reads that do not map uniquely to the genome.
- 73. There was strong consensus amongst respondents that the software algorithms used to assemble NGS sequence reads for *de novo* and re-sequencing purposes are immature, with developments lagging considerably behind the sequencing technological platforms. There is currently a mixed economy and 'cottage industry' of software, with researchers utilising, and continuously developing, numerous different algorithms, and often showing considerable loyalty to a familiar software platform. However, it is not clear which of the available methodologies represent the best approach for particular NGS assembly tasks, and respondents suggested that investment is required in the critical evaluation of the currently available methodologies. Following evaluation to identify those algorithms of broad applicability support should be provided for their maintenance and sustainability.
- 74. It was noted that a majority of algorithms use 'de Bruijn' graphs to assemble genome sequence relatively quickly and efficiently, but this does not fully utilise the information content of the dataset and relies heavily on heuristic approaches to resolve repeats. A research challenge linked to this observation is the development of probabilistic methods for resolving repeats and to extract the most likely assembly from the available data.
- 75. The challenges associated with repeats can be extreme in *de novo* sequencing and resequencing applications. In particular, it was highlighted that NGS approaches do not yet allow the highly repetitious sub-telomeric regions of eukaryotic genomes to be resolved. Unlike for Sanger sequencing, sub-telomeric DNA is well represented within raw NGS data outputs, but assembly of the sequence data remains extremely difficult, and greatly exacerbated by the short individual read lengths. Assembly of sub-telomeric regions therefore continues to represent a considerable technical challenge; indeed, one respondent noted that sequencing of the sub-telomeric regions has only been completed for a single genome, a task requiring a concerted research effort.
- 76. The consultation identified a consensus amongst respondents that the performance of sequence assembly algorithms depends strongly on the particular application and that tools need to be optimised for different systems (e.g. DNA from different species or different sequence read lengths). Furthermore, a significant challenge articulated by some respondents was that, for some of the key applications described under term of reference 1 (e.g. transcriptomics, large genomes and metagenomics), NGS algorithm development is currently highly deficient and is severely restricting developments.
- 77. The Expert Group recognised that well managed software development platforms are needed to provide a focus for the development of NGS software algorithms that have

generic use, harnessing existing distributed efforts, avoiding duplication, identifying robust solutions, 'hardening' these solutions and supporting maintenance and on-going improvement. The Collaborative Computational Projects (CCP) provide a model which could be appropriate for the development of NGS software solutions. Centralised facilities such as TGAC also have a role to play in coordinating and sustaining software development efforts. Support for the initiation of this type of activity could be provided via the Tools and Resources Development Fund and the Bioinformatics and Biological Resources Fund.

Conclusion 10: There is a shortage of suitable software for sequence assembly of NGS data and in some areas (e.g. metagenomics) it is highly deficient.

Recommendation 9: BBSRC should encourage proposals to fill the gap in sequence assembly algorithms via the Tools and Resources Development Fund and the Bioinformatics and Biological Resources Fund.

#### Data storage and transfer

- 78. In addition to challenges around sequence assembly software, there was a broad consensus that further logistical and technical challenges abound in the effective storage and transfer of NGS data. This consensus extends to the need for standardisation of formats and quality markers both for primary NGS data and the processed outputs (FASTQ files) to ensure that future added value can be obtained from archived material.
- 79. Several respondents noted that archiving of primary image and intensity data is prohibitively expensive due to their very large size (>TB), and that real-time analysis of primary data removes the need to store it. The emerging consensus seems to be that only processed FASTQ files, (~200GB) should be archived. However, this has yet to be standardized across the field, and ongoing step changes in storage capacity and cost reduction could lead to a revision of this situation in the foreseeable future.
- 80. There was acknowledgement that downstream analysis pipelines generate large volumes of data, including intermediate forms. It may be necessary to archive these data to comply with research data sharing policies and to underpinning publications or to ensure added value can be obtained from future research.
- 81. Movement of large quantities of data between sequencing facilities and users' laboratories currently represents a considerable logistical challenge. Many files are too large to move easily or quickly over the internet or via standard portable media. Currently the simplest solution seems to be to transfer files using a portable hard-drive. A clear picture of how this situation is expected to develop (improve or possibly worsen with further NGS and concurrent IT infrastructural advances) did not emerge from the consultation responses, and the situation may require ongoing monitoring.
- 82. The areas of NGS and IT data storage are both fast moving and it is difficult to predict the outcome of the current challenge that NGS data presents. It could, for example, simply be overcome by commercial developments in data compression. That said, the related area of sequence interrogation and analysis might require 'data-intensive' high performance computing (HPC) solutions. National capability in HPC is currently coordinated across several Research Councils. Issues of NGS data transfer are likely to be tackled together with other data-intensive science areas (e.g. meteorology and astronomy) and sit within the much bigger issue of the IT network infrastructure in the university sector. Multiple stakeholders are active here, including the Funding Councils and the Joint Information Systems Committee (JISC), as well as, the seven Research

Councils. It is recommended that BBSRC represent the current needs of the bioscience research communities in discussions on computational infrastructures (HPC, IT networking).

Conclusion 11: The massive expansion of sequence data being produced by NGS technologies is presenting challenges for commodity storage and data transfer. While storage is generally dealt with locally, transfer of data relies on the HEI network infrastructure.

Recommendation 10: BBSRC should ensure that the growing needs of the bioscience research community (including NGS data storage, interrogation and transfer issues) are considered in the multi-stakeholder discussions on e-infrastructures and IT networking.

#### Data Analysis, Data Interpretation, and the Skills Gap

- 83. There was a strong consensus amongst the consultation respondents that a major bottleneck of critical importance to the widespread adoption of NGS is the extraction of maximum biological knowledge from the data. The outstanding needs in data analysis and interpretation span research, skills and training, and infrastructures. They include bespoke algorithm development for analysis and reanalysis of the data; improving the skills of bioscience researchers so they can design the best NGS experiments and interrogate the resulting data, and integrate it with existing NGS and non-NGS data; and the development of more user friendly software for research driven by biological questions.
- 84. Several respondents commented that extraction of biological understanding would not be served by generic algorithm development (as is the case for sequence assembly), but by the design and application of bespoke algorithms by experienced and able researchers in computational genomics. It was recognised that a sea change in statistical and computational approaches to genomics data was on the horizon, where researchers will need to compute correlations across multiple massive datasets to derive the effects of particular genomic, epigenetic or environmental changes. As one respondent described for transcriptomics, "so many more transcripts will have to be considered in the analysis including small RNAs and splice variants. Subtle changes in splice variants can have a profound biological outcomes and NGS gives access to this information". There was a strong message in the consultation that there had been insufficient investment in the area and significant efforts were required to deal with the on-going shortage in bioinformatics expertise.
- 85. The Expert Group considered that within the UK, outside of the EBI and the Sanger Institute, there was a severe lack of capacity in bioinformatics and computational genomics. This was confirmed by the consultation responses, which almost unanimously identified shortages in bioinformatics, data analysis, and statistics together with the need to upgrade researchers' skills. As one respondent put it "The UK invented the current market leader in this technology, but we are not the world leader in its use".
- 86. The Expert Group agreed that focused effort is required to move from the current situation to a future with a well established, widespread computationally proficient bioscience research community. Members identified a number of mechanisms that could be adopted in order to boost capacity and fill the skills gap:

- Short courses / workshops: extensive hands-on workshop style training in the available tools to help disseminate expertise. Principally aimed at postgraduate and postdoctoral researchers;
- PhD training: to deal with the on-going, serious shortage of bioinformaticians and provide the next generation of appropriately skilled individuals;
- Fellowships: early stage and mid-career fellowships to provide leadership and build capacity by establishing research groups;
- Short-term access to computational expertise: Provide high-level informatics expertise into bioscience research labs to work on innovative problems for a defined period (e.g. 3-6 months);
- Discipline hopping: engage computational scientists, statisticians, mathematicians, physicists and other disciplines with expertise in working with large datasets to work data handling in short pilot / taster projects handling NGS outputs;
- Grants: No NGS grant should be funded that does not include an appropriate level of bioinformatics support.
- 87. One of the features of a centralised facility is that it acts as a hub for expertise and training. TGAC is well placed to deliver some of the mechanisms above (e.g. short courses, a 'home base' for bioinformaticians providing the short-term expertise), as well as provide a stimulating research environment for postgraduate students and research fellows.
- 88. Finally, on user-friendly software, respondents noted that current software is predominantly Linux-based and command-line driven, its operation requiring specialist computational skills. The development of web-based tools and graphical interfaces clearly represents an important next step towards lowering the technical skills barrier to software operation. If NGS hardware is to be effectively exploited by researchers in a distributed operating model, the lack of user-friendly interfaces would seem to be a barrier to localised use of software.

Conclusion 12: There is a major gap in NGS data analysis and data interpretation, which is exacerbated by a lack of capacity in computational genomics and bioinformatics.

Recommendation 11: BBSRC should, working with other partners as appropriate, develop a programme of activities to boost capacity and capability in bioinformatics and computational genomics, utilising the mechanisms identified in paragraph 86 and drawing on the expertise and experience at the centralised facilities (e.g. TGAC).

#### Infrastructures for data sharing and analysis

- 89. There was recognition amongst respondents that NGS will place considerable demands on community infrastructures for data sharing, which centre around the database and software provision of the EBI (EU) and National Center for Biotechnology Information (NCBI; North America). EBI hosts the EMBL Nucleotide Sequence Database (also known as EMBL-Bank), which constitutes Europe's primary nucleotide sequence resource, and represents the likely primary destination (along with the NCBI's short-read archive) of NGS outputs for public research access. The huge likely increase in submissions to this database brings considerable logistical and resourcing challenges for hosting, cataloguing and quality control of an expected data deluge.
- 90. Centres of critical mass and community focus for bioinformatics, such as EBI, are also at

the forefront of research efforts to produce software for the annotation and analysis of genome sequences (e.g. Ensembl). Similar pressures abound in bioinformatics tool development, particularly to allow annotation and analysis of new data types being facilitated by NGS (e.g. metagenomics, cellular dynamics).

91. The expected huge increases in 'omic data outputs from NGS approaches has, in part. informed the case for ELIXIR, a proposed, multi-nationally funded, >£200M upgrade to the EU bioinformatics infrastructure to be built around the existing hub of EMBL infrastructure at the EBI. BBSRC has already invested £10M (August 2009) to allow a dramatic increase in the EBI's data storage and handling capacity as a first step in developing the 'next generation' bioinformatics infrastructure. This initial investment and the full business case for ELIXIR have been made in cognisance of the advent of a step change in high throughput bioscience research, including NGS approaches, and of the Research Councils' related strategic investments in high throughput biology. These include BBSRC's six flagship system biology centres (circa £40M) and TGAC (circa £14M), and MRC's four High-Throughput Sequencing Hubs (£9M). The ELIXIR project is in a preparatory phase that is already funded. Maturation of the project is subject to successfully obtaining funding to an appropriate level from European nations, including a multi-funder bid<sup>5</sup> to the UK Large Facilities Capital Fund. If substantively supported, ELIXIR will provide: an infrastructure for integration of biological data, software tools and services throughout and beyond Europe; support for other European infrastructures in biomedical and environmental research; and services for the research community, including training and standards development. ELIXIR, if the project matures successfully, therefore has the potential to substantially provide the European infrastructure required to underpin high throughput biology, and NGS approaches therein. The Expert Group welcomed ELIXIR and strongly endorsed BBSRC's proactive engagement in the preparatory phase programme. However, in the event that ELIXIR is not substantially taken forward to maturity, these resourcing needs will remain largely unmet, and will need to be funded through other models, with huge challenges therein. Challenges also exist in ensuring that bioinformatics infrastructure is able to keep pace with the demands of researchers in advance of the outcomes of any substantive ELIXIR investments.

Conclusion 13: The ELIXIR project, if substantively funded, should go a long way to ensuring that bioscience researchers have access to appropriate bioinformatics infrastructure to enable the management and sharing of genome sequence datasets. However, considerable problems will arise in providing appropriate infrastructure if ELIXIR investments do not substantively materialise.

Recommendation 12: BBSRC should continue in its strong support for ELIXIR as demonstrated by proactive engagement in the preparatory phase and the recent financial commitment to upgrade the EBI's data infrastructure.

.

<sup>&</sup>lt;sup>5</sup> BBSRC, MRC, NERC and the Wellcome Trust.

#### **SUMMARY OF CONCLUSIONS AND RECOMMENDATIONS**

#### Conclusions

Conclusion 1: Next Generation Sequencing technologies are considered to offer several advantages (technological, cost and speed) over the Sanger methodology. Driven by huge market potential, improvements (2G) and new approaches (3G) are subject to intense commercial development. Collaboration between industry and academia on pre-commercial prototypes would be mutually beneficial.

Conclusion 2: NGS is having a significant impact across the biosciences. Current applications include *de novo* sequencing, re-sequencing, epigenetics and metagenomics. Over the next 5 years, NGS approaches are expected to become more widespread and well established and BBSRC has a role to play in facilitating their adoption by supporting 'taster sessions' and pilot studies.

Conclusion 3: Key areas of research and development that will benefit from NGS include food security (e.g. plant and animal breeding, disease diagnostics and management), industrial biotechnology (e.g. drug discovery, biopharmaceutical developments) and medicine (personalised medicine, disease diagnostics and management). The technology will have both economic and societal impact and the use of data arising from some potential applications (e.g. personalised medicine) raises significant ethical and societal issues.

Conclusion 4: There is currently an appropriate balance between the supply of NGS equipment and the demands of the BBSRC user community. However, this is a dynamic situation and one that could quickly change; technology availability will need to keep pace with demand.

Conclusion 5: There is a 'mixed economy' of NGS provision in the UK. The Research Councils and other funders have established centralised facilities whilst HEI's have set up local services. A significant proportion of NGS technology currently lies outside of the centralised facilities. Coordination is required to establish best practice, facilitate knowledge exchange and, importantly, to derive maximum value from the investments made.

Conclusion 6: TGAC will boost UK capacity for sequencing plant, animal and non-medical microbial genomes. Clarification of the range of activities that TGAC will provide to the research community is required. A TGAC technical assessment process should be adopted, wherever relevant, across BBSRC funding mechanisms.

Conclusion 7: The roll-out of NGS raises a number of funding issues: equipment cost and rapid improvement in machine specification makes it difficult to determine the 'right time to buy'; democratisation will continue and requests to fund machines and services outside TGAC are anticipated; and there is a need for greater transparency in the requests made to BBSRC to obtain funding.

Conclusion 8: The NGS data production pipeline is currently at an early stage of development. There is evidence that experimental design may be influenced by cost, while production bioinformatics platforms are in a state of flux and there is a serious gap in data analysis and interpretation capabilities. Large-scale infrastructures appear to be managing the current level of data submissions and are planning for future expansion.

Conclusion 9: Primary challenges for NGS chemical and molecular biology methodologies are centred on the effective parallelisation of sample preparation, and increasing sensitivities to allow reduced sample sizes. Addressing these challenges could increase the range of applications and lead to cost savings.

Conclusion 10: There is a shortage of suitable software for sequence assembly of NGS data and in some areas (e.g. metagenomics) it is highly deficient.

Conclusion 11: The massive expansion of sequence data being produced by NGS technologies is presenting challenges for commodity storage and data transfer. While storage is generally dealt with locally, transfer of data relies on the HEI network infrastructure.

Conclusion 12: There is a major gap in NGS data analysis and data interpretation, which is exacerbated by a lack of capacity in computational genomics and bioinformatics.

Conclusion 13: The ELIXIR project, if substantively funded, should go a long way to ensuring that bioscience researchers have access to appropriate bioinformatics infrastructure to enable the management and sharing of genome sequence datasets. However, considerable problems will arise in providing appropriate infrastructure if ELIXIR investments do not substantively materialise.

#### Recommendations

**Recommendation 1:** BBSRC should encourage collaboration and partnership between leading sequencing facilities (e.g. TGAC) and NGS equipment manufactures to facilitate refinement of, and early-stage access to, the very latest technologies.

**Recommendation 2:** BBSRC should facilitate the widespread adoption of NGS by supporting 'taster sessions' and pilot projects that will enable roll-out of the technology across the key areas of application (e.g. re-sequencing, gene expression, epigenetics, metagenomics) within its remit.

**Recommendation 3:** To ensure that economic and societal impact is derived from the application of NGS approaches, BBSRC should promote knowledge exchange between academic researchers at the forefront of developing and utilising the new technologies, and stakeholders in industry and government.

**Recommendation 4:** BBSRC should monitor the availability and use of NGS technology, in order to maintain an awareness of the balance between supply and demand.

**Recommendation 5:** BBSRC (working with other partners where appropriate) should facilitate the coordination of UK NGS activities, to establish best practice, facilitate knowledge exchange and, importantly, to derive maximum value from the investments made.

**Recommendation 6:** BBSRC should: (a) work with TGAC to clarify the range of activities that it will provide to the bioscience user community, and; (b) Adopt a technical assessment process for TGAC wherever relevant, across BBSRC funding mechanisms.

**Recommendation 7:** BBSRC should ask applicants to justify (i.e. on the basis of cost and quality) the choice of sequencing facility as a part of the research grant application process.

**Recommendation 8:** BBSRC should encourage responsive mode proposals in NGS chemical and molecular biology methodologies.

**Recommendation 9:** BBSRC should encourage proposals to fill the gap in sequence assembly algorithms via the Tools and Resources Development Fund and the Bioinformatics and Biological Resources Fund.

**Recommendation 10:** BBSRC should ensure that the growing needs of the bioscience research community (including NGS data storage, interrogation and transfer issues) are considered in the multi-stakeholder discussions on e-infrastructures and IT networking.

**Recommendation 11:** BBSRC should, working with other partners as appropriate, develop a programme of activities to boost capacity and capability in bioinformatics and computational genomics, utilising the mechanisms identified in paragraph 86 and drawing on the expertise and experience at the centralised facilities (e.g. TGAC).

**Recommendation 12:** BBSRC should continue in its strong support for ELIXIR as demonstrated by proactive engagement in the preparatory phase and the recent financial commitment to upgrade the EBI's data infrastructure.

#### **BBSRC**

February 2010

Annex 1

#### **BBSRC MISSION**

To promote and support, by any means, high-quality basic, strategic and applied research and related postgraduate training relating to the understanding and exploitation of biological systems.

To advance knowledge and technology (including the promotion and support of the exploitation of research outcomes), and provide trained scientists and engineers, which meet the needs of users and beneficiaries (including the agriculture, bioprocessing, chemical, food, healthcare, pharmaceutical and other biotechnological related industries), thereby contributing to the economic competitiveness of the United Kingdom and the quality of life.

In relation to the Council's activities, and as the Council may see fit, to:

- Generate public awareness
- Communicate research outcomes
- Encourage public engagement and dialogue
- Disseminate knowledge

#### MEMBERSHIP OF THE EXPERT REVIEW PANEL

#### **Members**

Name	Institution		
Professor Ottoline Leyser FRS (Chair)	University of York		
Professor Shankar Balabsubramanian	University of Cambridge		
Professor Richard Baldock	MRC Human Genetics Unit Edinburgh		
Dr Ewan Birney	EMBL European Bioinformatics Institute		
Dr Mike Dawson	Novacta Biosystems Limited		
Professor Keith Edwards	University of Bristol		
Professor Peter Kille	University of Cardiff		
Dr Jon Slate	University of Sheffield		
Dr Tony Smith	Independent Consultant		
Professor Claire Wathes	Royal Veterinary College		



#### NEXT GENERATION SEQUENCING REVIEW

#### RESEARCH COMMUNITY CONSULTATION QUESTIONNAIRE

#### **DEADLINE FOR RESPONSES: 9 AUGUST 2009**

#### **BACKGROUND**

In November 2008, BBSRC's Strategy Advisory Board agreed that a review of the current and future impact of next generation sequencing should be undertaken. The terms of reference for the review are:

Review developments in NGS and its potential impact on biological research (animals, plants, non-medical microbes).

Establish a view of the current availability of the NGS equipment in the UK and associated data storage and management arrangements.

Identify bottlenecks in NGS approaches (i.e. what issues need to be tackled to ensure that the full benefit of existing and future NGS approaches are realized) including (a) chemistry methodologies, molecular biology methodologies and informatics tools; and (b) bioinformatics infrastructures (software and hardware).

Provide advice to Strategy Advisory Board on activities and investment that may be required to ensure appropriate adoption of NGS approaches and data collection, analysis and sharing.

An Expert Group has been established to undertake the review under the chairmanship of Professor Ottoline Leyser FRS (University of York; member of BBSRC's Strategy Advisory Board).

Consultation with the research community is an important part of the review process and BBSRC is keen to obtain your views on next generation sequencing. The questionnaire is framed around each of the terms of reference above. Although the questionnaire will be returned by e-mail, no personal information will be entered on the questionnaire. The completed version of the questionnaire should be emailed to: <a href="mailto:nextgen@bbsrc.ac.uk">nextgen@bbsrc.ac.uk</a> on or before **9 August 2009**.

BBSRC thanks you in advance for taking the time to complete this questionnaire.

For further information about the review or this questionnaire, contact:

Dr Sophia Abbasi

Tel: 01793 413 027; <a href="mailto:sophia.abbasi@bbsrc.ac.uk">sophia.abbasi@bbsrc.ac.uk</a>

Contact Name:	
Contact Email:	
If you are responding	

on behalf of a society
or organisation, please
• .
provide details here.

#### General

**G1** Please briefly describe your areas of research.

### ToR1: Review developments in NGS and its potential impact on biological research (animals, plants, non-medical microbes).

- **1.1** What novel sequencing technologies do you expect will become available in the next 3-5 years?
- **1.2** What do you consider to be the main areas of application of NGS approaches to date? What new application areas do you think will arise in the next 3-5 years in your research area(s) and other areas of biological research?
- **1.3** What new scientific opportunities and new ways of working become possible with the adoption of NGS approaches (i.e. what are the key biological questions that could be addressed with these new technologies?)
- **1.4** What impact (i.e. scientific, economic, societal) do you think will arise from the new technologies?

### ToR2: Establish a view of the current availability of the NGS equipment in the UK, its use and associated data storage and management arrangements.

Do you use NGS technology?

If YES, please describe

- (a) the facility
- (b) how it is funded and maintained
- (c) access arrangements (i.e. how access is obtained direct, or service; how long does it take to get access etc)
- (d) what you use it for
- (e) challenges (e.g. logistical, technical, financial) in adopting the new approaches
- (f) the infrastructure and arrangements in place for handling the NGS data-sets
- (g) when and how data generated from the NGS facility is shared with the wider research community?

If **NO**, please state why not (e.g., lack of availability, see no use for NGS in current research area; too expensive for current research; unable to handle the resulting data; waiting for

further technological developments).
What constitutes robust experimental design for studies using NGS? General comments and specific examples would be very useful.
What constitutes robust data management for studies using NGS?
Are standards or minimum information systems under development or in place for the reporting of data, information and knowledge generated using NGS?
How does UK NGS provision compare to that available in other countries. What could the UK learn from international NGS provision?
ToR3: Identify bottlenecks in NGS approaches (i.e. what issues need to be tackled to ensure that the full benefit of existing and future NGS approaches are realised), including (a) chemistry methodologies, molecular biology methodologies and informatics tools; and (b) bioinformatics infrastructures (software and hardware).
3.1 What are the research challenges in chemical and molecular biology methodologies?
<b>3.2</b> What research challenges need to be tackled to expand the use of NGS approaches in functional genomics?
3.3 What are the challenges in (a) generating the NGS data sets (b) storing the data arising from NGS; and (c) Extracting information and knowledge from the data sets (inc. assembly algorithms)?  Please describe the challenges together with thoughts on potential solutions.
<ul><li>(a) generating the NGS data sets</li><li>(b) storing the data arising from NGS; and</li><li>(c) Extracting information and knowledge from the data sets (inc. assembly algorithms)?</li></ul>
<ul> <li>(a) generating the NGS data sets</li> <li>(b) storing the data arising from NGS; and</li> <li>(c) Extracting information and knowledge from the data sets (inc. assembly algorithms)?     Please describe the challenges together with thoughts on potential solutions.     </li> <li>3.4 Are there skills gaps that impede the full utilisation of NGS approaches? If so, please</li> </ul>

Additional comments in line with the terms of reference of the review will also be welcome.

Thank you for providing your views.

#### RESEARCH COMMUNITY CONSULTATION RESPONDENTS

Responses were received from individuals affiliated with the following organisations:

- Aberystwyth University (Institute of Biological, Environmental and Rural Sciences)
- Cancer Research UK (Cambridge Research Institute)
- Food & Environment Research Agency
- Illumina Ltd
- Imperial College of Science, Technology and Medicine
- BBSRC Institute of Food Research
- BBSRC John Innes Centre (3 responses)
- New Zealand Blood Service
- Oxford Nanopore Ltd
- Roche Diagnostics Ltd
- Scottish Crop Research Institute
- University of Birmingham (2 responses)
- University of Bristol
- University of Cambridge (4 responses)
- University of Edinburgh (5 responses, 2 from the Roslin Institute)
- University of Exeter
- University of Leeds
- University of Liverpool (4 responses)
- University of Nottingham (3 responses),
- · University of Oxford
- University of Manchester
- University of Sheffield
- University of Southampton (2 responses)
- University of Warwick (2 responses)
- Wellcome Trust Sanger Institute

#### **BBSRC GRANTS RELEVANT TO NEXT GENERATION SEQUENCING**

Type of Award	Gabriel Reference	Award Title	Award- holding Institution	PI and co- applicant(s)	Start Date	End Date	Total Award Value
Grant	BB/C507902/1	Transcription factor interactions during mesendoderm formation in Xenopus	University of Cambridge	Smith J Argasinska J	20041101	20071031	£312,323
Grant	BB/C519370/1	Sequencing the tomato genome: a reference genome for the Solanaceae	The Wellcome Trust Sanger Institute	Rogers J	20050501	20081130	£401,833
Grant	BB/E004091/1	A computational platform for the high-throughput identification of short RNAs and their targets in plants	University of East Anglia	Moulton V Baulcombe D Dalmay T	20070226	20100225	£250,576
Grant	BB/E006981/1	SIROtyping : siRNA and miRNA profiles of tomato	University of East Anglia	Baulcombe D Dalmay T	20070301	20070831	£43,555
Grant	BB/E006981/2	SIROtyping : siRNA and miRNA profiles of tomato	University of Cambridge	Baulcombe D Dalmay T	20071101	20100430	£639,953
Grant	BB/E007228/1	SIROtyping : siRNA and miRNA profiles of tomato	University of Warwick	Manning K	20061101	20091031	£17,077

Grant	BB/E018130/1	Populations genetics and genomics of ovine nematode parasites and their application to study the molecular basis of anthelmintic resistance.	The Wellcome Trust Sanger Institute	Parkhill J Berriman M	20070901	20100831	£144,442
Grant	BB/E020909/1	Changes in gene expression during sex chromosome evolution in the dioecious plant Silene latifolia	University of Edinburgh	Charlesworth D Bergero R	20071101	20101031	£342,310
Grant	BB/E021832/1	Impact of mutations in the target-encoding CYP51 gene in Mycosphaerella graminicola populations developing resistance to triazole fungicides.	Swansea University	Kelly SL Kelly DE	20080201	20110131	£291,703
Grant	BB/E02257X/1	Impact of mutations in the target-encoding CYP51 gene in Mycosphaerella graminicola populations developing resistance to	Rothamsted Research (RR)	Fraaije B Cools HJ Lucas JA	20080225	20110524	£225,115

		triazole fungicides					
Grant	BB/E023568/1	From rice to orphan crops: robust, high	John Innes Centre (JIC)	Griffiths S Snape JW	20070601	20080930	£69,019
		throughput genetic markers for all the grasses.					
Grant	BB/E023576/1	From rice to orphan crops: robust, high throughput genetic markers for all the grasses	Institute of Grassland and Environmental Research (IGER)	King IP King J	20070901	20080331	£7,768
Grant	BB/E023576/2	From rice to orphan crops: robust, high throughput genetic markers for all the grasses	Aberystwyth University	King IP King J	20080401	20080831	£2,199
Grant	BB/E024866/1	Genome-wide analysis of short RNAs as modulators in dehydration stress tolerance using tolerant and genetic model systems	University of East Anglia	Dalmay T	20070515	20100514	£313,466
Grant	BB/E025013/1	Massively Parallel Sanger Sequencing by Duplex Melting	University of Oxford	Mir KU	20070621	20080620	£99,964

Grant	BB/F00334X/1	The genome sequence for the potato cyst nematode Globodera pallida and its utilisation for improved control	The Wellcome Trust Sanger Institute	Parkhill J Berriman M	20080301	20110228	£943,385
Grant	BB/F007523/1	Large scale mapping of wheat transcripts and simultaneously defining the A, B and D genome contribution to the hexaploid transcriptome	University of Bristol	Edwards KJ	20080204	20120203	£465,031
Grant	BB/F007981/1	Nutrition and early life programming: exploring epigenetic mechanisms	Newcastle University	Relton CL	20080401	20110331	£400,180
Grant	BB/F009313/1	Sequencing the transcriptome of Kalanchoe fedtschenkoi: a model for Crassulacean acid metabolism, embryogenic plantlet formation and the Saxifragales	University of Liverpool	Hartwell J Hall N	20080506	20110505	£638,745

Grant	BB/F015917/1	Whole genome SNP panels for genotyping experimental chicken lines- a vital BBSRC resource	Institute for Animal Health (IAH)	Kaiser P	20080512	20091111	£98,263
Grant	BB/F016190/1	Using Solexa/Illumina methods to investigate plant pathogen variation and transcriptome	University of East Anglia	Jones JDG	20080313	20090612	£105,942
Grant	BB/F019394/1	A genome-wide association study of non-pathological cognitive ageing	University of Edinburgh	Deary IJ Porteous D Tenesa A	20080901	20100831	£698,047
Grant	BB/F019793/1	Resequencing Arabidopsis thaliana	EMBL - European Bioinformatics Institute	Birney E	20090201	20110131	£182,817
Grant	BB/F022441/1	A genome-wide association study of non-pathological cognitive ageing	The University of Manchester	Pendleton N Horan M Ke X Ollier W Payton A Pickles AR	20080901	20100831	£449,722
Grant	BB/F022697/1	Resequencing Arabidopsis thaliana	University of Oxford	Mott R Kover P	20080901	20100831	£486,546

Grant	BB/F02293X/1	Epigenetic events in micronutrient programming during early development	University of Cambridge	Dunger D Affara NA Constancia M Owens S Prentice A	20090102	20120101	£543,228
Grant	BB/G000093/1	pubmed2ensembl: a resource for linking biological literature to genome sequences	The University of Manchester	Bergman CM Nenadic G	20090223	20100222	£99,333
Grant	BB/G000298/1	Exploitation of new bacteriophages for generic strain engineering methods and functional genomic analysis of diverse bacteria	University of Cambridge	Salmond GPC	20081115	20100214	£100,806
Grant	BB/G000573/1	Tools for motif recognition in fungi	University of Liverpool	Wong P Caddick MX Hall N Rigden DJ	20090423	20100722	£103,066
Grant	BB/G002975/1	Dissection of a novel molecular pathway involved in seasonal timing in a melatonintarget tissue using an experimental and systems-level approach	University of Edinburgh	Burt DW	20090105	20130104	£782,679

Grant	BB/G003033/1	Dissection of a novel molecular pathway involved in seasonal timing in a melatonintarget tissue using an experimental and systems-level approach.	The University of Manchester	Loudon A Davis JRE	20081101	20121031	£789,800
Grant	BB/G006199/1	Characterisation of tomato short RNAs involved in fruit development	University of Nottingham	Seymour G	20090105	20120104	£5,034
Grant	BB/G00661X/1	Genomic analysis of complex speciation in Heliconius	University of Edinburgh	Blaxter ML	20090701	20120630	£104,485
Grant	BB/G006903/1	Genomic analysis of complex speciation in Heliconius	University College London	Mallet J Dasmahapatra KK	20090901	20120831	£513,501
Grant	BB/G00711X/1	A systems-level study of the role of epigenetics in mediating in utero environmental influences on genome function and transgenerational effects	Queen Mary, University of London	Rakyan V	20090501	20120430	£310,300
Grant	BB/G007721/1	Characterisation of tomato short RNAs involved in fruit development	University of Warwick	Manning K	20090126	20120125	£296,808

Grant	BB/G008078/1	Characterisation of tomato short RNAs involved in fruit development	University of East Anglia	Dalmay T Moulton V Szittya G	20090105	20120504	£618,703
Grant	BB/G008337/1	Chemical Mapping of G- Quadruplexes in the Genome	University of Cambridge	Balasubramanian S	20090521	20120520	£320,061
Grant	BB/G008841/1	Genomic analysis of complex speciation in Heliconius	University of Cambridge	Jiggins CD	20090401	20120331	£36,461
Grant	BB/G012016/1	Evolution of multidrug resistance in Salmonella enterica serovar Typhimurium as a result of biocide exposure.	University of Birmingham	Webber MA Pallen MJ Piddock LJV	20090713	20120712	£439,433
Grant	BB/G012075/1	A systems biology based approach to functionally annotate and analyse the genome of the fish pathogenic oomycete Saprolegnia parasitica	University of Aberdeen	van West P Secombes CJ	20090720	20120719	£324,433
Grant	BB/G012865/1	Mining the allohexaploid wheat genome for useful sequence polymorphisms	University of Bristol	Edwards KJ Barker G	20090501	20110430	£208,216

Grant	BB/G013004/1	Mining the allohexaploid wheat genome for useful sequence polymorphisms	University of Liverpool	Hall N Hall AJW	20090501	20110430	£1,068,856
Grant	BB/G013985/1	Mining the allohexaploid wheat genome for useful sequence polymorphisms	John Innes Centre (JIC)	Bevan MW	20090313	20110312	£294,997
Grant	BB/G015678/1	The Stat3 pathway in self-renewal, pluripotency, and the germline	University of Cambridge	Smith A Bertone P Nichols J	20090601	20130531	£1,211,904
Studentship	BB/G017883/1	Metagenomics of microbial communities on human skin	University of Liverpool	Hall N James AG McCarthy A	20091001	20130930	£74,410
Studentship	BB/G017980/1	Metagenomic analysis of bacterial populations associated with canine oral health through pyrosequencing	University of Liverpool	Allison HE Horsburgh MJ Marshall-Jones ZV	20091001	20130930	£74,410
Grant	BB/G020418/1	Integrated transcriptome and genetic analysis of early events determining tissue susceptibility in the Claviceps purpurea - wheat interaction	National Inst of Agricultural Botany	O'Sullivan D Barker G Bayles RA Gordon A Haseloff JP	20090601	20130531	£913,583

Grant	BB/G021821/1	Stem cell screening of human nutrient- gene interactions at the epigenetic level	University of Nottingham	Young LE Denning C	20090801	20120731	£1,011,291
Grant	BB/G02197X/1	A pipeline of resistance genes to Phytophthora infestans from wild Solanum species and their accelerated isolation using Illumina sequencing methods	University of East Anglia	Jones JDG	20090601	20120531	£761,302
Grant	BB/G02264X/1	EMBOSS: European Molecular Biology Open Software Suite	EMBL - European Bioinformatics Institute	Rice PM	20090501	20111231	£759,311
Grant	BB/G024650/1	Assessing Illumina and Velvet for sequencing and assembling a wheat chromosome arm.	John Innes Centre (JIC)	Bevan MW	20090601	20100531	£71,686
Grant	BB/G024715/1	Assessing Illumina and Velvet for sequencing a wheat chromosome arm	EMBL - European Bioinformatics Institute	Birney E	20091001	20100930	£39,582

Grant	BB/G024928/1	Investigating the role of short RNAs on wood formation, cambium development and adaptation of poplar tree (POPsRNA)	University of East Anglia	Dalmay T	20090401	20120331	£340,341
Grant	BB/G024952/1	Associative expression and systems analysis of complex traits in oilseed rape / canola - ASSYST (PRR-CROPP)	John Innes Centre (JIC)	Bancroft I	20090901	20120831	£362,429
Grant	BB/G024979/1	Plant Alternative Splicing and Abiotic Stress (PASAS)	University of Dundee	Brown JWS	20090701	20120630	£400,308
Grant	BB/G024995/1	BLOOM-NET	University of Leeds	Davies B	20090401	20120331	£288,059
Grant	BB/H00436X/1	Optimization of wheat and oilseed rape straw coproducts for bioalcohol production	Institute of Food Research (IFR)	Waldron K Faulds CB	20091001	20120930	£308,092
Grant	BBS/B/01839	Towards association genetics in wheat via a very high throughput genotyping method	University of Bristol		20040906	20070905	£247,279

Grant	BBS/B/02401	High-throughput reverse genetics to determine legume gene function	University of East Anglia	Jones JDG Wang TL	20040601	20070930	£326,238
Project	BBS/E/C/00004911	The genome sequence for the potato cyst nematode Globodera pallida and its utilisation for improved control		Kerry BR	20080101	20110630	
Project	BBS/E/C/00004981	Soil protection and remediation		McGrath S.P.	20080401	20130331	
Project	BBS/E/I/00001169	BBSRC Studentship: The role of pathogenicity islands in Salmonella gallinarum identified y genome sequencing in virulence for birds		Stevens MP	20041001	20070930	
Project	BBS/E/I/00001367	Whole genome SNP panels for genotyping experimental chicken lines - a vital BBSRC resource		Kaiser P Fife MS	20080512	20091111	
Project	BBS/E/J/00000584	Plant genomics and growth control		Bevan MW	19860101	29990101	

Project	BBS/E/J/000CA316	From rice to orphan crops: robust, high throughput genetic markers for all the grasses		Griffiths S Snape JW	20070601	20080930	
Project	BBS/E/J/000CA337	An integrated informatics and resources platform for Reverse Genetics in dicots (RevGenUK)		Wang TL Clarke JH Oldroyd G Ostergaard L Trick M	20080401	20120331	
Project	BBS/E/J/000CA375	Mining the allohexaploid wheat genome for useful sequence polymorphisms		Bevan MW	20090313	20110312	
Grant	E20156	Evaluation of multi-dimensional CE and multi- dimensional CE- MS for complex bioanalysis	King's College London	Smith NW	20040519	20070518	£194,764
Grant	REI18431	Population, genomic and evolutionary biology of complex systems	University of Edinburgh	Blaxter ML Andolfatto P Charlesworth B Charlesworth D Earl CRA Keightley P Maizels R Smith A	20030401	20080401	£157,076

## The Genome Analysis Centre Capacity and Capability Challenge

TGAC's Capacity and Capability Challenge (CCC) offers UK researchers the opportunity to engage with the Centre's new sequencing and bioinformatics facilities through its early access research programme. Commencing January 2010 and running for 12-18 months, the CCC will deliver a series of innovative projects addressing not only biological research problems but also technical challenges to sequencing and associated informatics.

Collaborative proposals involving academic and industrial partners are welcome under the CCC, as are links to established sequencing and bioinformatics centres worldwide. Sequencing funded by existing BBSRC research grants can also be directed through the programme.

NOTE: Companies wishing to secure sequencing directly from TGAC on a fee-for-service basis should not apply to the programme, but should contact the TGAC Business Development Director to discuss requirements.

This document contains the following information:

- Introduction to TGAC
- Scope of the Call
- Eligibility
- How to apply:
- Application form
- o Technical feasibility statement
- Assessment procedure and prioritisation
- Timetable
- Contacts

### Introduction

The Genome Analysis Centre (TGAC) is a UK resource for large-scale sequencing of plants, animals and microbes. TGAC is a significant investment in bioscience research infrastructure led by the Biotechnology and Biological Sciences Research Council in a funding partnership with the East of England Development Agency and Norfolk Local Authorities (Norfolk County Council, South Norfolk District Council, Norwich City Council and the Greater Norwich Development Partnership) TGAC provides much needed expertise and resources in genomics for these organisms including:

- High throughput and "next generation" sequencing;
- New technology platforms;
- Bioinformatics;
- Innovation, enterprise and skills activities.

The Centre complements the <u>Wellcome Trust Sanger Institute</u> and smaller research oriented sequencing hubs funded by the <u>Medical Research Council</u> and the <u>Natural Environment Research Council</u> and works in partnership with the <u>European Bioinformatics Institute</u>.

TGAC was launched in July 2009 and is being established in three phases:

- Start Up: equipment set-up, recruitment of personnel, establishment of the Scientific Advisory Board, establishment of a programme of prototype sequencing and related bioinformatics projects (the CCC call), and development of a plan for business engagement;
- Scale Up: Expansion of capacity for sequence generation and analysis; engaging in large and small scale genomics projects in community partnerships (national and international);
- Fully Operational: delivering large and small-scale genomics projects, deploying further new sequencing technologies, development of new data collection and distribution strategies, development of bioinformatics research and service support activities.

The TGAC Capacity and Capability Challenge will operate during the start-up and scale up phases (start date: Jan 2010; duration 12-18 months). During scale-up, CCC activities will be rolled into the full strategic programme of sequencing projects. The full programme will be established through stakeholder consultation, under the direction of the Scientific Advisory Board, chaired by Professor Sir John Sulston FRS.

### Scope of the Call

TGAC's Capacity and Capability Challenge (CCC) is an opportunity for UK researchers to engage with the Centre's new sequencing and bioinformatics facilities in its early access programme. Commencing January 2010 and running for 12-18 months, the CCC will deliver a programme of pilot studies that address not only biological research problems but also technical challenges to sequencing and associated informatics.

Collaborative proposals involving academic and industrial partners are welcome under the CCC, as are links to established sequencing and bioinformatics centres worldwide.

Sequencing funded by existing BBSRC research grants can also be directed through the programme. Companies wishing to secure sequencing directly from TGAC on a fee-for-service basis should not apply through this route, but should contact the <u>TGAC Business</u> <u>Development Director</u> to discuss requirements.

The CCC call will remain open until the full TGAC sequencing programme begins and it is expected that CCC projects ongoing at that time will be rolled into the full sequencing programme. It is also anticipated that many CCC projects may lead to long term and large-scale collaborative sequencing projects.

BBSRC (Swindon Office) is administering the CCC on behalf of TGAC to support the Centre during the 'start-up' phase.

When considering the submission of proposals researchers should be aware of the International Sequencing Consortium database that lists completed, on-going and future planned large scale sequencing efforts. See <a href="http://www.intlgenome.org/viewDatabase.cfm">http://www.intlgenome.org/viewDatabase.cfm</a>

### **Eligibility**

Higher Education Institutions (HEIs) and Research Council Institutes (RCIs) for which the BBSRC has established a long term involvement as major funder as part of the national research base may apply, together with BBSRC approved Independent Research Organisations. Further details on eligibility are provided in the BBSRC Grants Guide. Collaborative proposals between academic and industrial partners are welcomed under the CCC.

### **How to Apply**

You should submit a TGAC CCC proposal using the <u>application form</u>. All proposals need to be accompanied by a completed TGAC CCC <u>technical assessment form</u>, completed by the applicant and signed by the TGAC point of contact.

Where appropriate, and as a part of completion of the technical assessment form, applicants should demonstrate the capacity to provide TGAC with sample material of a high quality suitable for analysis. A small amount of funding can be requested to cover the staff time and consumables associated with sample preparation if funding to provide this resource if not already in place.

### **Assessment procedure and prioritisation**

The Scientific Advisory Board, chaired by Professor Sir John Sulston FRS, will review and prioritise proposals to the CCC programme, to assemble a pilot programme that will support TGAC to position itself as a world-class sequencing and bioinformatics centre. In reviewing and prioritising the proposals the Scientific Advisory Board will include the following criteria in its consideration:

- Strategic relevance of the proposed work to the funding partners.
- Feasibility of the proposed work, in line with the evolving capability and capacity of TGAC.
- Capacity to take forward the output received from TGAC.
- Cost effectiveness of any sample preparation request
- Availability of biological material of sufficient quality

The CCC programme will generate data that must be shared in compliance with BBSRC's <u>Data Sharing Policy</u> and in line with the well-developed practices that exist in the international genome sequencing community.

#### **Timetable**

It is anticipated that the first set of CCC projects will start in January 2010.

The CCC call will operate to a series of batching dates set at intervals to facilitate efficient review of proposals. The first set of batching dates is:

- 15 December 2009
- 28 February 2010
- 30 April 2010

### **Contacts**

Please contact TGAC (<u>TGAC.CCC@bbsrc.ac.uk</u>) to discuss any enquiries relating to the CCC Programme. For more general enquiries please email <u>TGAC.enquiries@bbsrc.ac.uk</u>.



## TGAC Capacity and Capability Challenge Programme

Part A - To be completed by the applicant				
2. Proposed investigators (Add or remove rows as required)				
Role	Name	Organisation	Division or	
Dairenianal			Department	
Principal Investigator				
Co-Investigator				
Co-Investigator				
Co-Investigator				
	tigator's Contac	t Details (full postal, teleph	none and email)	
4. Project Partner	s			
	ails of partnering	laboration with industry or organisations, and their pr		

- **4. Summary of proposed work** (<2 pages, to address points a f below)
- a. Aims;
- b. Proposed methodologies and timescales (including any timescale limitations);
- c. What support is required from TGAC (i.e. sequencing/informatics)?
- d. Do you have funding to support the work (in full, or preparatory phase and/or subsequent data analyses)?
- e. technical information (how long will it take to produce samples of appropriate quantity and quality? Will your samples present any biological hazards requiring particular precautions to be taken?):
- f. How will the output received from TGAC be analysed and released? How will this activity be resourced? Do you wish someone from your research centre to work with TGAC staff on a particular aspect of analysis?

Summary of Proposed Work:				

# Part B – Technical Assessment (To be completed by TGAC) 1. Date received by TGAC 2. Overall Assessment Is the application appropriate for access to TGAC resources during the CCC Programme? Yes/No 3. Feasibility Brief commentary on technical feasibility of the proposed work, including: • appropriateness of scale, selected methodologies, expected timescales; • identified risks; matching of work demands to TGAC capabilities and resources to deliver desired outputs. Name: Position:

This form, with Part A completed, should be submitted to TGAC at <a href="mailto:TGAC.CCC@bbsrc.ac.uk">TGAC.CCC@bbsrc.ac.uk</a>. Please allow a minimum of 10 working days for consideration and return of the assessment to you.

Date: