Professor Dame Janet Thornton
Director
T + 44 (0)1223 494648
F + 44 (0)1223 494496
thornton@ebi.ac.uk

Professor Dame Sally Davies
Chief Medical Officer
Department of Health
Richmond House, Room 123b
79 Whitehall
London
SW1A 2NS

21 March 2013

Dear Sally,

Re: 100,000 Whole Genomes Project
Letter from the Chief Medical Officer's Data Working Group

In January this year, you invited me to chair the above working group to consider the necessary standards, infrastructure and expertise that would be required, ideally building on existing platforms to provide the necessary data for clinicians and researchers. You also asked our working group to take into account the Government's wider programme of activities to support Big Data and to provide platforms for both public and commercial researchers.

The advice and opinion on how an integrated, interoperable data management framework can support the aims of the 100,000 Whole Genome project is presented below, with detail in the accompanying annexes. However, it must be stressed that this is a first look at the many and varied issues involved. There will be a need for continued work going forward to refine and formulate more detailed specifications for commissioning of services. The Data Working Group would be happy to provide what further advice it can to support you in this important initiative.

The Group looked at three specific areas:

1. Data Infrastructure and Flow,
2. Data Specification and Standards

3. Training and Workforce Development

Detailed summaries from the group for each area are in the annexes to this letter. An appendix, with specific exemplars of how sequencing data are currently handled within the NHS for Cancer, Rare diseases, HIV & TB is attached.

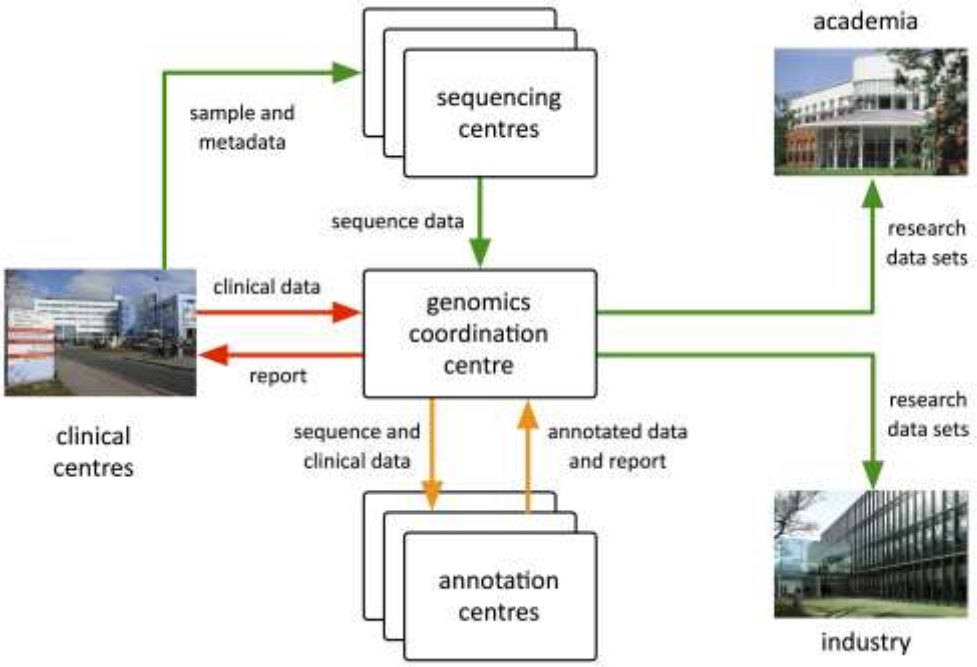
Whatever data are stored, it is the opinion of the group that data accumulated during the sequencing of the 100,000 whole genomes will provide a powerful resource for interpreting both existing and future cohorts. The development of appropriate structures and systems for sharing relevant data, whilst maintaining privacy, must be a key factor of any future consideration on the development of the informatics infrastructure surrounding this project.

1. Data Infrastructure and Flow

Handling the data and getting right the analysis, annotation and interpretation of sequence data is arguably the most essential part of this initiative. If this is not right from the outset, it could undermine the success we all want to see. The quality of the data, how it is stored and shared, common meta-data and other standards will need to be drafted and agreed. As the programme progresses, this will need to be systematically reviewed to ensure the standards and procedures remain fit for purpose.

Connectivity and interoperability need to be primary goals if we are to produce datasets and databases providing the richest possible information on how disease develops and how to treat it. In order to harness the economic potential, how data are shared, in what format and with whom, will need to be an integral part of policy development and implementation. To achieve this, we need to ensure that there is confidence in process and outcomes: for academia, research, industry and, most importantly, patients and the public.

This final point is vital. That is why the Data Working Group recommends the operational framework and architecture shown below:



At the heart of this framework is a Genomics Coordinating Centre (GCC), which should be established within the NHS network to benefit from the security and confidence this brand brings. It would provide a tangible and co-ordinated biomedical informatics function, starting with the 100,000 genomes data, providing governance and a platform for life sciences innovation and enterprise. The role of the GCC would include:

- Delivering a clear and transparent operational and governance framework
- Setting standards for data encoding, translation and interoperability frameworks
- Setting standards for biomedical informatics education and training
- Interfacing to clinical data and clinicians
- Co-ordinating the provision of genetic variant interpretation services in the NHS to improve patient care
- Developing a UK-wide database of whole genome sequences
- Providing high-end bioinformatics services to the NHS
- Supporting life sciences enterprise by providing access to data and working closely with industry to identify commercial opportunities

It would be accountable to the relevant governance bodies and, ultimately, Parliament. With time, this could evolve to take a wider role and provide trusted and credible leadership in all areas of biomedical informatics.

The architecture we propose would allow for multiple providers for each part of the delivery pathway. Clear, detailed specifications of the data flows will need to be agreed, even where sequencing and annotation are performed within the same organisation. Provision will need to be made for the storage and transmission of large data files. Within this architecture, there is a need for routine quality control of the sequencing data and annotations and appropriate validation.

A more general approach is required for informatics here, to the extent that there is little prospect for the re-use of existing software. However, the experience gained may be extremely valuable.

The control centre will manage the release of suitably-abstracted clinical and genomic data sets for research purposes. This will be an important part of its remit, ensuring that information rich data can be accessed in a secure environment based on an agreed governance framework. This will hopefully ensure patient and public confidence in the system. Although this architecture will serve both human and pathogen sequencing, the sequencing and annotation centres may be different, depending on technical and clinical imperatives (eg turn-around time and genome size) and current practices in the different parts of the NHS. However, the role of the GCC in setting common standard, specifying services, monitoring activity, and controlling access would remain important in each case.

The GCC will work closely with the Health & Social Care Information Centre and the Clinical Practice Research Datalink as appropriate, but it is important to stress that neither of these organisations has the right expertise, or appropriately trained staff, to perform the GCC role. The CR UK Stratified Medicines Programme, the UK10k project, the DDD (Deciphering Developmental Disorders) project and the HIVrdb serve as useful exemplars on how data infrastructure and flow can be developed. Some information on these has been included in the Annexes.

2. Data Standards and Specifications

Specific standards must be in place for clinical data, genome sequences and their annotations. The GCC will establish and maintain these standards, taking account of input and feedback from prospective users. The data standards consist of two types, the raw reads from the sequencing services and the individual variations from the reference genome – the variant calls.

The raw data is approximately 100Gb per whole genome. Cancer genomes require sequencing both the normal genome (at high coverage) and the tumour genome at even higher coverage to provide good quality information, which will occupy around 300 GB for raw data.

For most clinical purposes, it is likely that only variant data will be needed. However, we believe that raw data reads will need to be stored and made accessible; as a valuable data set for research and help to improve the technology to interpret them, which will be vital for accurate clinical interpretation The technical standards to compress and store the raw data should be addressed urgently.

The raw sequence data could be managed as a separate resource, or stored by the sequencing companies for recall by the control centre. However, at the scale of around ~100,000 genomes, predominantly cancer sequencing, the expected total disk requirement is substantial, between 10 ~ 20 Petabytes of disk. As well as the disk cost itself, there are important engineering components to execute well at this scale. It is important to realise that we are *not* recommending that storage of raw data is considered to be a clinical standard in the future, but at the current state understanding of genomic data processing, it is prudent to store the raw data for this cohort.

Variants come in different types, ranging from single base pair changes (SNVs), short insertions and deletions, through to larger copy number variants and complex rearrangements (Structural Variants). Accurately calling each type of variant will be important in many clinical applications. The data volume of variant calls is in the order of 2 GB per genome.

Access to variant calls and associated clinical data would be available under controlled access, with a Data Access Committee, appointed and overseen by the GCC, establishing principles and making decisions. Knowledge abstracted from these data would be published through open access data resources.

We recommend the following features for the genomic data formats.

- Whilst there remains uncertainty about genomic information processing, raw data from whole genome sequencing should ideally be stored for several years from generation. However, we recognise that some practical limits on long term raw data storage may be necessary.
- Called variants for whole genome or exome cancer and germline variants are stored as Variant Call Format (VCF), including uncertainty information and negative information (being able to differentiate between a lack of data and reference genotypes).
- Variant Call information for pathogens is tailored to the genome and analysis process of that sequencing, but wherever possible existing standards should be used.

- Variant calls from raw data should be treated as part of the experimental process, and should achieve high quality levels for both sensitivity and specificity. A provider must describe their calling process and assess their false positive and false negative rates for the variants they are providing.
- Meta-data on both sample identification/tracking numbers are present in both formats, and a process flow of identifying potential tracking errors, in particular between institutions is created to understand the likely points of failure. Other meta-data such as reference genome build version and analytical software used should be provided.

All these decisions should be reviewed regularly to ensure that the remain relevant. A detailed paper describing recommended standards for sequences is at Annex 1b.

3. Training and development of a bioinformatics workforce

The Group also considered the strategic workforce planning issues arising from the predicted expansion in information flows. This work goes beyond the 100,000 whole genome initiative and has a multi-agency stakeholder constituency, including the Department of Health, NHS Commissioning Board, Public Health England and Health Education England, as well as the Higher Education and Biotechnology sectors.

The rapid expansion in genome sequence and other related information will require partnerships in new innovative models of delivery between health, academia and the private sector. It could also require the introduction of new types of regulated clinical professionals notably in medicine (genomicist) and in healthcare science (Clinical Scientist in Clinical Bioinformatics). There may also be some increase in the current medical and scientific specialist genetics, genomics and specialist public health microbiology workforce. The long lead in time, and complexities to train and develop some specialist professional groups, require a strategic approach to workforce planning, education and training programme development to ensure any rapid expansion is thought through carefully and managed systematically. There should be a more general approach to improve knowledge and skills in the opportunities and implications of genomic data and bioinformatics across the healthcare workforce.

Whole genome analysis should be applied across the NHS only at the point when clinical interpretation of a genome can be provided, otherwise existing models of health care would be subverted. This is why workforce development is an essential first step in harnessing potential healthcare benefits. Work is already underway to develop training programmes, both in the NHS and the research community, and these should be further developed to ensure the speedy and continued adoption of genomics and bioinformatics training. Programmes from both fields must complement and synergistically enhance each other; promoting commonality rather than

difference, and providing a platform for the necessary research and development capacity and capability building.

The training and workforce development programmes should:

- Provide the skills base and improve the capacity to effectively manage and interpret genome-scale data in clinical genetics, e-health record research, and clinical bioinformatics communities.
- Build upon existing provision to establish new targeted and specific education and training and research programmes at postgraduate level (Masters and doctoral) linked to accompanying workplace training supporting new and defined career pathways together with a recognition of the need for continual CPD and updating workshops to provide ongoing knowledge and skill development in this rapidly changing field.
- Promote cross-fertilisation between academic and health sectors through joint workshops focusing on genome bioinformatics, e-health record research and variant interpretation as well exploring the shared training opportunities including with the private sector.

To achieve this, the group believes the following actions should be undertaken, with commitments obtained from the education and training system to ensure responsiveness in 2013-14:

- A multi-agency strategy should be developed to implement programmes that will increase training in biomedical genomics based on robust workforce modelling to provide evidenced-based, comprehensive information on and analysis of future demands.
- Re-train and develop in the order of 20% of the current 150 clinical geneticist population as practitioners of a new specialism, clinical genomics.
- Develop a new strand of medical speciality training to increase the number of clinical genomicists and clinicians (e.g. pathologists and oncologists) with genomics expertise.
- Embed the healthcare science workforce's clinical bioinformatics speciality and career pathway, including a new Masters level pre-registration training programme and a higher specialist scientific training programme.
- Continue to recruit and develop clinical scientists in (a) genetics with an enhanced skill and knowledge base in biomedical genomics at all levels of the career pathway to expand and continue their work alongside clinical geneticists and genomicists, and (b) in specialist microbiology to support the expansion in the genomics of infectious diseases.
- Develop the necessary Continuing Professional Development (CPD) programmes to support the education and training of medical, nursing, science, pharmacy and managerial staff in this domain (required numbers remain to be estimated).

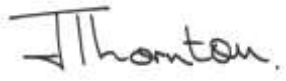
- As an urgent priority, increase postgraduate and post-doctoral training capacity and capability in order to 'train the trainers'.
- Promote cross-fertilisation between academic, health, research and industry sectors through collaborative working on genome bioinformatics, e-health record research and variant interpretation.
- Activity and funding should be concentrated into a smaller number of education and training centres in both NHS and HE sectors to provide the necessary short term traction.
- Pursue joint training and research opportunities in collaboration with Health eResearch Centre initiatives such as cross sector workforce training placements, continuing professional development at the bio-health informatics interface, sub-specialty medical training, and health e-Research support.

Detailed papers on education and training and workforce development are provided at Annex 2 and 2a.

In closing, I would like to thank the members of the Data Working Group for all their excellent support and hard work. We all believe that the 100,000 genomes initiative has excellent potential to build upon existing UK leadership in genomes, research and the life sciences industry. It will be an important opportunity for the NHS to develop world-class sequencing and informatics capabilities to provide improved diagnostic services and support the development of new therapies. Having the NHS leading the process will help ensure that the 'trusted brand' can ensure patient and public confidence.

I hope that this brief report of a complex landscape is of some help in taking the initiative forward. I look forward to hearing from you if there is any more that I, or the whole group, can offer.

Yours sincerely



Professor Dame Janet Thornton DBE, FRS
Chair, CMO Data Working Group

Appendix

Current Practices in NHS today according to specialist area

Cancer

Briefing notes on Cancer Research UK Stratified Medicine Programme

Aims of the Programme

During Phase 1 (2011–2013)

- Demonstrate model of standardised, cost-effective, routine somatic mutation testing in the NHS.
- Develop informatics and technology solutions to support the annotation of clinical samples
- Test for at least 5-10 mutations including existing well-validated biomarkers linked to treatment.
- Collect tissues, clinical data and mutation results on 9,000 samples from lung, prostate, ovary, colorectal, breast and malignant melanoma. Cases originate from 7 clinical hubs and 3 technology hubs receive samples for sequencing and return reports.

End of Phase 1/commencement of Proposed Phase 2

- Integrate lessons learned from Phase One into broader clinical practice in NHS.
- Adopt use of newer technologies including next generation sequencing.
- Include other cancer types and applications in other disease areas.
- Broaden scope and utility of datasets captured

A data driven architecture for Stratified Medicine

Phase 1 of the Stratified Medicine Programme (SMP) has piloted molecular diagnosis of cancer in the NHS. An Informatics Solution was required both to evaluate the programme and to demonstrate integration of molecular diagnostics into NHS data flows.

Methodology

The underpinning principle of the informatics solution for SMP has been to build on existing NHS computer and data systems. The data flows are based on known standards for health care data and where possible built on funded work for national data analysis and reporting. The Programme has established pseudonymised exchange of data between clinical and technology hubs. This was based on an XML message generated from the hospital system and sent to laboratories. When the test is completed results are populated into the laboratory databases. A results message is then generated automatically and

returned to the hospital. Structured extracts flow from the clinical hubs to Eastern Cancer Registry and Information Centre (ECRIC).

The dataset produced for use in the programme has been adapted from the latest version of the NCIN's national cancer outcomes and services dataset (COSD), which is due to replace the current National Cancer dataset in England and to be implemented in 2012. In Scotland there is no systematic collection of all cancer-related patient data in a disease-specific manner and there is a separate cancer dataset for Wales (All Wales Core Cancer Minimum Reporting Requirement, current v5.0). The data items selected map to the NHS data dictionary and the data items in the final section have been created specifically for use in the Stratified Medicine Programme (SMP).
For the purposes of SMP, data is hosted within the Eastern Cancer Registry Information Centre (ECRIC) and stored securely behind an NHS firewall. In the future, it is intended that data will be released in an anonymised manner to approved researchers for research purposes, subject to the necessary ethical approval having been granted by NRES or a local research ethics committee. A data access committee for the programme is currently being created and will oversee the anonymisation process.

Learning Points from Implementation

One of the key principles of the Programme was to evaluate how far along the data quality spectrum we could get with routinely recorded clinical records and what level of proactive data management would be required to make it fit for research purposes.

Linkage of the data

The clinical data collected by the SMP reflects the patient pathway with a multitude of one-to-many relationships particularly with patients experiencing many investigations, procedures, treatments and different gene tests. The SMP data model is patient-centric, to reflect the research study component and to the future needs of molecular pathology. However, ECRIC maintains a classic cancer registry data model that is based around the tumour and this has led to issues linking the data during its integration into the ECRIC database. The linkage issues and inconsistent use of the concatenated SMP unique identifier have resulted in particular problems with one-to-many relationships generating multiple duplicate records for each patient, making it difficult to link molecular results to tumour samples for patients with multiple pathology specimens and linking treatment records for a small number of patients with more than one primary cancer (e.g. co-existing primary colorectal and lung cancer).

Data Quality

Issues with data completeness have mainly been focused on particular data items with histopathology data, co-morbidity scoring system data e.g. ACE-27 and performance status proving particularly challenging. This is mainly due to a lack of inter-operability between information technology systems used to generate and store histopathology reports and other parts of the electronic patient record. SMP has worked with the clinical hubs to improve the

completeness of the data. In many cases, the solutions have required manual intervention by data managers even for those systems that are automated for other data feeds. In some cases it has highlighted areas that will need clarification for the national rollout of the Cancer Outcomes and Services Dataset in January 2013 and this feedback is being provided to NCIN. For example, in some sites all patients ACE-27 data are incorrectly coded as 'severe' based on the diagnosis of cancer. However the aim of this scoring is to assess overall fitness for treatment and should be scored based on comorbidities.
Cross-border data issues between England, Scotland and Wales have been encountered, resulting in problems sourcing some demographic details although suitably de-identified surrogates (such as year of birth rather than full date of birth) have been agreed.

Improving Data Quality

ECRIC have completed a series of detailed analyses of the data that have identified areas where there is a mixture between use of codes versus free text and in some cases invalid values (despite the specified attributes for the majority of the data items). This feedback has been provided to the hubs in a series of data reports that highlight any inconsistent data items at the level of individual records. This has also allowed us to identify areas where further guidance may be required. We have also been able to identify where mapping between coding systems may be required. For example, in Wales, READ codes are in use rather than SNOMED CT and in the NHS in England several different editions of SNOMED are in use at the other sites and these will be mapped against the core programme dataset.

Rare Diseases

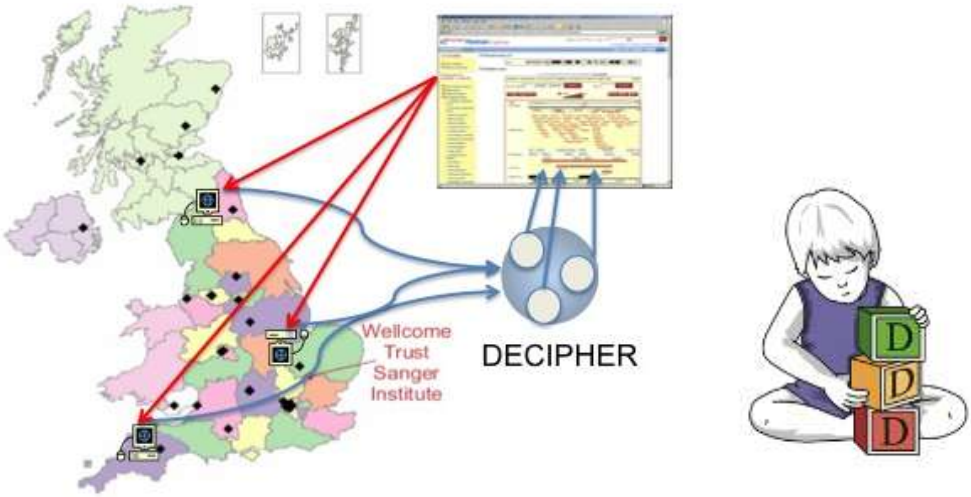
Delivering 10k exomes in the NHS - insights from the Deciphering Developmental Disorders (DDD) study that may help develop strategy for rare diseases within the 100k genomes challenge

Key points
DDD/DECIPHER provides an end-to-end prototype for scaled NHS-based multi-centre recruitment of patients, identification of diagnostic genetic variants from genome-wide data and dissemination of diagnostic results to referring clinicians. This model is critically dependent on sufficient resourcing of bioinformatics and collection of structured high-quality clinical information, including family history and phenotypic observations.
Interpretation of variation within coding regions is challenging but tractable, both from clinical (i.e. diagnostic) and research (e.g. novel causal gene discovery) perspectives, however, clinical and research interpretation of non-coding variation is much less tractable.
The rate of change of knowledge about genes known to cause rare diseases is very rapid; to maximise diagnostic utility this necessitates a system of genetic data storage capable of iterative reporting against an actively curated and updated database of known causal genes

DECIPHER (http://decipher.sanger.ac.uk)
DECIPHER is a web-based database (established in 2004) to correlate patient phenotype (clinical features) with genomic variants to aid interpretation of genetic data to support clinical care and research. DECIPHER contains records on >21,000 patients with rare copy-number variants contributed by a network of ~200 centres worldwide, and is used by all 23 NHS Regional Genetics Services. It is partitioned into 'private' password protected domains and publicly accessible linked-anonymized information that is served globally with patient consent. From March 2013, DECIPHER will integrate sequence variation and incorporate additional visualization and interpretation tools to enable integrated investigation of all major classes of pathogenic variant.

Filtering patient variants findings against an evolving knowledge-base of common variants and disease-causing variants – Plausibly pathogenic variants identified by whole genome or exome studies need to be rapidly filtered against common variants and known-disease causing genes/variants to enable rapid identification of potentially diagnostic findings. DECIPHER incorporates up-to-date resources of known common variants and known pathogenic genes; and also provides a catalogue of individual instances of phenotype-variant combinations to facilitate identification of overlapping causal variants in patients with similar clinical features, and catalyse novel disease-gene or disease-variant associations.

Deciphering Developmental Disorders (www.ddduk.org)
DDD is a UK-wide collaboration between the 23 Regional Genetics services and the WT Sanger Institute to undertake exon-array and exome sequencing analysis for 12,000 patients with severe undiagnosed developmental disorders (severe or extreme phenotypes present from birth or early childhood). The Health Innovation Challenge Fund funds the DDD study (2010-2015). Currently >4000 patients with diverse phenotypes have been recruited by >150 Clinical Genetics consultants (>95% of all such NHS consultants). The phenotype and clinical data is entered by the local clinical team via a secure web-based module in DECIPHER and sample–tracking initiated by scanning a DNA collection tube. The link between the DECIPHER ID and patient data is held securely within the local NHS centre. Clinician's email addresses are included in the DDD record and when results are ready they are displayed in each centre's private DDD-DECIPHER domain and an email is sent to the responsible clinician with a link to the patient's result. Results can be viewed in their genomic context within DECIPHER and downloaded as a PDF for the patient record.



Developmental Disorders Gene to Phenotype (DDG2P) database – Within the DDD project, to support the identification of pathogenic variation in patients, we have established a clinician-curated gene panel of existing and newly reported genes causing developmental disorders. Genes in the panel are classified with information necessary to support computational filtering of candidate variants prior to manual review (e.g. by their mode of action eg. heterozygous/biallelic/hemizygous and by the consequence of the variant associated with the phenotype/disease), see below. This list is updated on a monthly basis by searching high impact journals for reports of new disease genes/variants relevant to developmental disorders. This approach enables new diagnostic findings to be returned rapidly to the clinician to aid patient management. Existing genomic data can be refiltered against the updated gene list periodically. Variants that are not prioritized by the DDG2P filter remain available for ongoing research.

Establishing specialty led networks for recruitment and dissemination of results
The DECIPHER/DDD model could be broadened to encompass other clinical specialties outside of clinical genetics by customizing the informatics modules that enable patient recruitment, rapid interpretation and dissemination of genomic results to NHS services for each medical specialty; with each specialty customising its own patient clinical data sheet for recording clinical information necessary for interpretation of genetic data and developing relevant gene

panel(s) for filtering variants to identify diagnostic results whilst enabling ongoing research on undiagnosed patients.
With the approval of the Academy of Royal Colleges and engagement of the academic divisions of national specialty groups it should be feasible to (i) rapidly identify patients with rare undiagnosed disorders of likely genetic origin and (ii) develop specialty specific clinical data sheets and dynamic gene panels for relevant phenotypes eg. peripheral neuropathy, ataxia, or immunodeficiency. This approach would enable rapid penetration of genomic technology in the NHS and maximize patient diagnosis across the diversity of rare diseases.

- Applying growing knowledge about genetics in real-time to improve diagnosis for NHS patients

- Interpretation of variants in a system well integrated with existing global resources for genomic variation

- Variants re-filtered against evolving gene panels enables patients to benefit from new discoveries

- Patients identified and recruited by clinicians through a distributed network enables accurate phenotyping and rapid return of diagnostic results to patients.

- Importance of recruiting patients and archiving DNA in regional centres/teaching hospitals

A secure web-based system for recruitment and dissemination of results facilitates equity of access for patients with rare disease across all regions of the UK

Rare diseases have high morbidity and mortality – some patients may die before a genomic diagnosis is available. Archived DNA with the requisite mutation(s) enables genetic testing to be offered to relatives.
Local DNA extraction and storage ensures that an archived sample is available locally for validation and as a positive control to facilitate future clinical testing of family members

Genetics specialists are available in regional centres/teaching hospitals to help with interpretation of genomic results and to discuss with clinicians the relevance of the genomic findings to the patient's phenotype/disease. They can also take forward cascade testing of relatives or discussion of reproductive options such as pre-implantation or prenatal diagnosis where appropriate.

A distributed approach will drive genomic technology into mainstream medical specialties whilst utilising genetic networks to share expertise regarding interpretation and clinical significance of genomic variants and to facilitate appropriate integration of genomics into the patient pathway.

References

Firth HV, Wright CF. The Deciphering Developmental Disorders (DDD) study. Dev Med Child Neurol 2011;53:702-03 PMID: 21679367
Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP. DECIPHER: DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources. Am J Hum Genet. 2009;84:524-33. PMID: 19344873
Swaminathan GJ, Bragin E, Chatzimichali EA, Corpas M, Bevan AP, Wright CF, Carter NP, Hurles ME, Firth HV. DECIPHER: Web-based, community resource for clinical interpretation of rare variants in developmental disorders. Hum Mol Genet. 2012 Oct 15;21(R1):R37-44. PMID: 22962312
HVF/MEH 29.1.13

Pathogens - HIV

Generation and reporting of sequence for HIV clinical utility

Samples – routinely obtained from HIV clinics. Referred to one of 15 England NHS virology labs with sequencing capability. 10,000 sequences generated /year in UK

Sequencing labs – NHS, capillary sequencing 1kb per sample (only 10 % of viral genome). UK network supports quality and methodology. Immediate interpretation into drug susceptibility predictions using web based algorithms eg http://hivdb.stanford.edu/

Reporting – direct report from virology labs back to clinician to assist in therapy decisions for individual patients. Turn around 1-2 weeks

Databases

UK HIV drug resistance database held at MRC Clinical Trials Unit. Formal governance structure involving all large clinics and all laboratories. Steering Committee approves all research proposals and use of data. See website: http://www.hivrdb.org/

Submission of sequences – laboratories (above) submit FASTA files of each sequence annually, with limited (pseud-anonymisation) patient identifiers

Linkage with phenotype

a) The UK Collaborative HIV Cohort study receives clinical data from the major HIV clinics in UK. Includes laboratory markers, and clinical details eg therapies. Also formal governance structure. These data are linked to sequence data at the MRC Clinical Trials Unit. See website: http://www.ukchic.org.uk/
b) HPA HIV surveillance data. All infections in UK have allied demographic data eg risk group. These data are linked to sequence data at the MRC Clinical Trials Unit

Annotation
All sequences can be interpreted for each analysis using a range of bioinformatic tools addressing identification of resistance mutations, virus subtype, polymorphic mixtures, phylodynamics

Outputs
a) Annual report from HPA including trends in drug resistance, transmitted drug resistance etc
b) Research- relates to genotype- phenotype relationships, and molecular epidemiology. Often link data with that of international collaborators to provide sufficient size.
c) Industry – partial funding from Industry, and provide industry with access to specific analyses

Current Funding

1. Sequence data. This is an NHS diagnostic service, within contracts with each laboratory. NHS cost around £200/sequence
2. UK HIV Drug Resistance Database and UK Collaborative HIV Cohort Study funded by MRC Programme Grant (UCL)
3. HPA. Commissioned by DH to produce annual HIV drug resistance reports.

Human Exome sequencing

Current proposal for NIHR Bioresource exome sequencing linked to clinical database. Patients actively recruited 2013-4. Exome sequencing 2014-5

Implications of moving to viral Whole Genome Sequencing in future

1. NHS labs would wish to move their sequencing capacity towards NGS, allowing continued real time reporting back to clinicians to guide therapy. Capital implications unclear.
2. However, urgent requirement for laboratory methodologies and web based assembly and interpretation tools to enable this to happen. Currently such details being developed by WTSI, with view to roll out.
3. Aim for current databases to assimilate NGS whole genomes, using current model.
4. However, essential for database to move to more secure (not grant based) funding

Pathogens – TB

Data for TB Services, current position

A total of 8,963 cases of tuberculosis were reported in the UK in 2011, with 5284 of those cases confirmed by culture of a range of clinical samples, including sputum and tissues. Whole genome sequencing can only be carried out on

cultures at present, although in the future, direct extraction and analysis of patient specimens may be suitable.
The majority of isolates of mycobacteria are cultured, (patient specimens, such as sputum, incubated in nutrient broth until growth is detected on a semi – automated system within high containment laboratories(CL3) to avoid spread of infection to staff) in more than 100 NHS laboratories in England, and sent from relevant geographical areas in the North, Midlands and South to the HPA reference laboratory in Newcastle, Birmingham or London for identification, susceptibility determination and molecular typing.
Identification, (as TB or non-TB mycobacterium species) is carried out by molecular techniques, using commercial systems based on detection of combinations of oligonucleotides, and reported to sender laboratories and clinicians in 24-48 hours.

Notification of cases to Public Health, contact tracing and treatment initiation are triggered by report of a culture identified as TB. Cases may also be treated and notified by clinicians, based on clinical findings.
Sensitivity of all new isolates of TB to anti tuberculous drugs is determined phenotypically, taking 7-14 days after the organism has grown, and reported to sender laboratories and clinicians. Molecular tests to detect genetic mutations that reliably detect resistance for some commonly used drugs are used regularly but not universally, when resistance is suspected on clinical grounds or from phenotypic tests, to shorten the time to recognition of drug resistance.

TB reference laboratories in England and Wales undertake epidemiological typing by 24 locus MIRU-VNTR on all new cases in England, generating a 24 digit profile from enumeration of tandem repeats at 24 loci in the genome. Profiles are reported locally to clinicians and public health teams, and to the HPA national database, compared against the database to define the phylogenetic clade with which it clusters, which often provides information on the likely geographic origin of the source strain, and numbered if clustered with other isolates with indistinguishable profiles.

Linkage between genome sequence and clinical data
All the information reported by laboratories to NHS service users is also reported centrally to HPA databases , held in Colindale, and contributes invaluable data for national surveillance.

Analysis of clusters is carried out locally, particularly in high incidence areas, and nationally , linking typing data with clinical, treatment outcome and risk factor information that clinicians and TB nurses enter on to the HPA Enhanced TB Surveillance database. This is database is available to NHS users to access information on patients under or transferring to their care, protected by access codes by the Health Protection Units, as well as for outbreak investigation, cluster analysis and cohort review

The current mechanism within the HPA to link at the national level clinical, epidemiological and molecular typing by 24 Locus MIRU VNTR could be adapted to the use of whole genome sequence data for tuberculosis.

Whole genome sequencing is not part of routine testing, and is currently a research tool only. A project in Oxford and Birmingham to implement microbial whole genome sequencing for individual patient care, including identification and sensitivity testing of isolates, local outbreak recognition and national surveillance, funded by the DH/Wellcome Trust Health Innovation Challenge Fund is one of several projects that will address this. Sequencing of *M. tuberculosis* and accumulation of bioinformatics expertise is also being performed in Cambridge (Wellcome Trust Sanger Institute).

Current Funding

1. Culture of specimens is funded by local NHS hospitals
2. Identification, sensitivity testing and molecular typing is funded by the HPA (since the 1960s). Molecular typing by MIRU VNTR costs approximately £50 per isolate, and WGS would cost a similar sum or less, it is estimated.
3. Typing and clinical databases (ETS) are nationally funded by the HPA, but larger urban centres with high rates of TB also have their own clinical databases. The largest of these, the London TB Register, is currently being linked to the national ETS database, to reduce duplication of effort.

Implications of moving to Whole Genome Sequencing in future.

The accuracy of whole genome sequence to predict or refute *M. tuberculosis* transmission is under evaluation by several research groups, and scrutiny of the published data will be required before genome analysis can be performed in routine clinical practice. Subsequent validation of its accuracy will be an important objective during the early phases of clinical use. As far as we are aware, there is no software available to undertake this analysis in an automated fashion, and analysis currently relies on an experienced bioinformatician. It will be essential to develop an automated interpretation pipeline that generates meaningful output to healthcare workers and is readily accessible to the NHS.

Whole genome sequence could provide an accurate method of bacterial identification, and would obviate the need for alternative identification methods. This approach is also under evaluation by research groups and is not yet in clinical use.

Molecular tests have been in routine use for several years that detect genetic mutations that are reliably associated with resistance. This forms the basis for a catalogue of gene mutations that could be detected by whole genome sequencing to predict phenotypic resistance. For several drugs, testing remains reliant on phenotypic susceptibility testing which, because of the slow growth of *M. tuberculosis,* take weeks or even months to complete.

Phenotypic tests are performed when there is a poor understanding of the genetic basis of resistance. There will be a continued need to validate the phenotypic effect of novel gene mutations. Furthermore, it will be more straightforward in the early days of using genome sequencing to be confident of resistance based on the presence of a known gene mutation, than susceptibility in the absence of known mutations. This is because resistance mechanisms may be highly complex and involve more than one gene. Overall, whole genome

sequencing could potentially simplify current testing protocols, and would be predicted to be cost saving – especially as the phenotypic relevance of gene mutations is specifically evaluated and collated.
Larger NHS labs in areas of high incidence of TB would wish to move their sequencing capacity towards NGS, allowing continued real time reporting back to clinicians to guide therapy. Capital implications unclear.
Laboratory methodologies and web based assembly and interpretation tools are needed to enable this to happen.

All participants should supply clinical and genomic data for national surveillance, outbreak detection and management to a secure database, which must be accessible locally and nationally (including the Devolved Nations) by clinicians and public health practitioners.

Grace Smith 13.3.13

Annex 1

Infrastructure Requirements for 100,000 Genomes

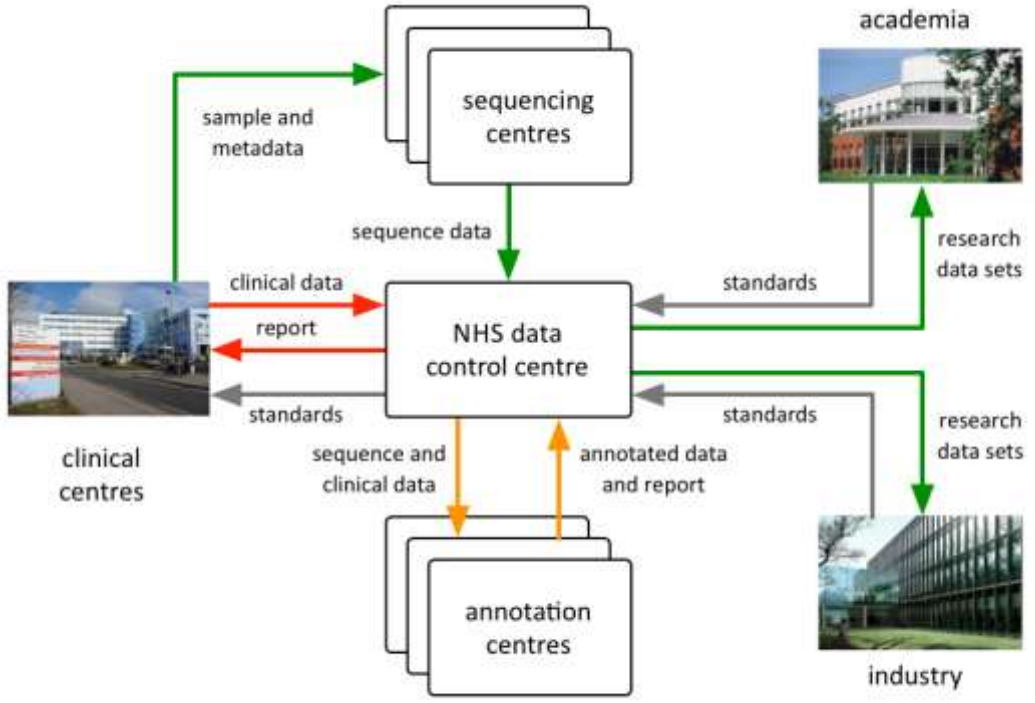
Working Group on Infrastructure (Jeff Barrett, James Brenton, Jim Davies, and Janet Thornton)

Headline Recommendation

A data control centre should be established, within the NHS network, with well-defined interfaces to sequencing and annotation services. The centre should hold the identifying mappings, define the data standards, and generate anonymised data sets. There will be specific informatics requirements upon clinical, sequencing, and annotation centres. Some provision will need to be made for the storage and transmission of large data files.

Architecture

The diagram below shows the proposed architecture. Red, orange, and green indicate data flows at different levels of identification. The "clinical data" flow from clinical centres to the control centre will include sample management metadata: in particular, an identifier representing the package sent for sequencing. The control centre will link clinical and sequence data, and send an abstracted version for annotation.



The architecture allows for multiple organisations performing each role. Clear, detailed specifications of the data flows will be required, even where sequencing and annotation are performed within the same organisation. The control centre will establish and maintain standards for clinical and genomic data, taking account of input and feedback from prospective users; these flows are shown in grey on the diagram.

As a minor variation upon the architecture, the sequence data (in particular, the raw data) could be managed as a separate resource. A similar variation may be required for a pathogen sequence data. The role of the centre in specifying services, monitoring activity, and controlling access would remain the same in each case.

Informatics

Standards: The initial data and metadata specifications should be established in advance. This can be done on the basis of previous study protocols, input from potential data users, and existing data standards. Some additional informatics support may be required for the collection of clinical data and sample tracking metadata. The value domain for every item must be clearly specified, together with the detailed procedure for collection. An agreed clinical terminology should be used where feasible.

Samples: The clinical data for each sample should be sent to the control centre before the sample is sent for sequencing. A unique identifier should be generated for the sample: the control centre will maintain the link between this and the identifiers used for the associated clinical data, sample metadata, sequence data, annotations, reports, and any version of the data produced for research purposes. Depending upon the study area, some data will need to be sent to the sequencing centre: for example, gender or cancer type.

Coping with change: The data and metadata specifications will need to be updated during the programme to reflect advances in sequencing technology, progress in scientific research, and feedback from data users. The informatics infrastructure should be designed to accommodate these updates, and to cope also with changes in clinical systems and procedures. In particular, the repository for clinical data should employ a logical records model will be needed.

Sequencing: Sequencing centres will need to process the data generated and deliver data in specified formats – for example, particular versions of the Variant Call Format – to the control centre, or to a separate NHS-controlled storage service. Raw sequence data should be archived: it need not be readily available, but should be retrievable for research or if it is deemed appropriate to repeat the processing.

Annotation: Annotation centres will need to perform both generic and disease-specific annotation, using public reference disease databases. Along with the annotations, they will deliver a clinical report summarising relevant variants. This report will be added to the records repository maintained by the control centre, and passed back to the associated clinical centre for inclusion in the patient record.

Control: The control centre will manage the release of suitably-abstracted clinical and genomic data sets for research purposes. It will hold patient and organisational data to support this, including consent. A Data Access Committee will need to be established to develop principles for academic and commercial use of the data and oversee applications for data access from researchers.

Exemplars: The CR UK Stratified Medicines Programme, the UK10k project, the DDD (Deciphering Developmental Disorders) project and the HIVrdb serve as useful exemplars, and some information on these has been included as an Appendix. A more general approach is required for informatics here, to the extent that there is little prospect for the re-use of existing software. However, the experience gained may be extremely valuable.

Information Centre: The Health and Social Care Information Centre (HSCIC) is charged with the management of NHS data as a publicly-owned asset for a number of purposes, including research. It should not be assumed that the HSCIC would be able to take on the role of the control centre: the standards will be driven by research needs, and many of the data items will fall outside its existing experience. The HSCIC is however well-positioned to assist with the definition of standards for clinical data and its transfer and storage.

Annex 2

Data Format Sub Group: Jeff Barrett, Ewan Birney, James Brenton, Peter Donnelly, Paul Flicek, Matt Hurles, Gil McVean and Simon Tavaré.

We expect the following flow of information for cancer and germline information

1. Production and alignment of "raw" reads. The data volume is around 100 GB per high coverage full germline genome. Cancer genomes require sequencing both the normal genome (at high coverage) and the tumour genome at even higher coverage to provide good quality information. A Normal/Tumour pair will occupy around 300 GB for raw data.

2. The calling of variants from raw data. Variants come in different types, ranging from single base pair changes (SNVs), short insertions and deletions, through to larger copy number variants and complex rearrangements (Structural Variants). Accurately calling each type of variant will be important in many clinical applications. With current state-of-the-art methods, the quality of the calling will differ between different classes of variant. In the case of cancer, this contrasts the normal genome with the cancer genome, making the process even more complex. The data volume of variant calls is in the order of 2 GB per genome.

3. The annotation and filtering of the variants to those of clinical relevance.

We expect the following flow of information for pathogen-related information

1. Raw data reads, being either Sanger or Next Generation Reads. Data size will vary, but be less than human genomes.

2. A variety of reports or analysis depending on the genome. In many cases, but not all, this will include assembly.

We recommend the following features for the genomic data formats.

1. As there is still uncertainty about genomic information processing (i.e., going from steps 1 to 2 above), we recommend that the raw data is stored for at least 10 years from generation. Current best practice is to use BAM moving to the compressed CRAM format over time. Lossy compression on quality information could be used, with appropriate consideration of its impact. These technical decisions should be reviewed regularly.

2. We recommend that called variants for whole genome or exome cancer and whole genome or exome germline variants are stored as Variant Call Format (VCF), including uncertainty information and negative information (being able to differentiate between a lack of data and reference genotypes). Current best practice is to use a VCF format, but there is diversity in how to handle negative information. A more precise

VCF definition would have to be created to be precise about both of these points, and again this technical decision should be reviewed regularly.

3. We recommend that call information for pathogens is tailored to the genome and analysis process of that sequencing, but wherever possible existing standards should be used. These technical decisions should be reviewed regularly.

4. We recommend that variant calls from raw data should be treated as part of the experimental process, and should achieve high quality levels for both sensitivity and specificity. A provider must describe their calling process and assess their false positive and false negative rates for the variants they are providing. Current best practice is that this should be >99% accuracy at known genetic variations, and <1% false positive rate for new mutations in germline samples; other variant types and other sample types (cancer, pathogen) will have their own quality standards.

5. We recommend that meta-data on both sample identification/tracking numbers are present in both formats, and a process flow of identifying potential tracking errors, in particular between institutions is created to understand the likely points of failure. Other meta data such as reference genome build version, and analytical software used should be provided, and these technical decisions should be reviewed regularly.

Currently the storage requirements of 100 GB (cost ~ £50) for a single high coverage full genome compares reasonably to data generation costs (currently around £3000-4000, hoped to drop to £1,000). Note that Cancer genomes require both a high coverage normal genome and an even higher coverage (approximately twice the coverage) of tumor genome.

However, at the scale of around ~100,000 genomes, predominantly cancer sequencing, the expected total disk requirement is substantial, between 10 ~ 20 Petabytes of disk. As well as the disk cost itself, there are important engineering components to execute well at this scale. It is important to realise that we are *not* recommending that storage of raw data is considered to be a clinical standard in the future, but at the current state understanding of genomic data processing, it is prudent to store the raw data for this cohort. Although the engineering to achieve accessible data storage at this scale is non trivial, it is clearly feasible.

Genomic Data format group would like to emphasise the following broader points

1. The data accumulated during the sequencing of this cohort will provide a powerful resource for interpreting both existing and future cohorts. The development of appropriate structures and systems for sharing relevant data while maintaining privacy need also be considered in the development of the informatics infrastructure surrounding this project.

2. Processing genomic data to provide genetic “calls” is still an evolving area of research, in particular in more complex variations (indels and structural variants) and in more complex scenarios (Cancer and Bacteria). This means that even small differences in calling behaviour can become complex confounders in other large scale analysis. This is why there is a need for raw data access for an appreciable period of time.

3. The analysis of genetic and genomic data has a number of complex error modes and artefacts (related to 1) and one needs to involve genetic and genomics analysts as well as more traditional statisticians and clinicians for the full utilisation of the data.

4. In large complex data flows there are often complex failure modes in the transfer and tracking of data items. Robust engineering practices for this scale of data flow, coupled with appropriate physical provision for network, storage and compute is required.

Annex 3

Training Needs: Report from the training sub-group

Group members: Andrew Devereau, Sue Hill, Pat Oakley, Colin Pavelin, Chris Ponting, Simon Tavaré, Janet Thornton

Individuals consulted: Sir John Tooke (UCL), Harry Hemingway (UCL), Angela Davies (Nowgen, Manchester), Helen Firth (Cambridge), Clare Turnbull (Royal Marsden), Val Davison (Birmingham), Angela Douglas (Liverpool).

Summary: Developing the Workforce

There are acute skills shortages in the workforce required to deliver the future benefits of genomic medicine[i]. Wide-ranging training initiatives will be required for medical and scientific staff who request and interpret genetic and genomic tests, and for clinical bioinformaticians and others who curate, annotate and analyse genome sequence data[ii,iii]. A preliminary Health Service workforce model (whose data and assumptions are described in Annexes 1 and 2) indicates, under certain conditions, that hundreds of individuals will need to be trained to meet shortfalls in the health service genomics workforce (by 2020: Clinical Geneticists: 56; Clinical Genomicists: 14; Clinical Scientists in Genetics & Bioinformatics: 289; Public Health Specialists: 28; and, Specialist Microbiologists: 19).

This increased training burden will fall on Universities and Research Institutes at a time when postgraduate and postdoctoral training in genetics, genomics and bioinformatics is already limited in both capacity and in capability and also on the NHS (and PHE) for clinically faced training where placement and training capacity is currently stretched . Although the academic workforce modelling has yet to be performed, it is likely that the number of suitably qualified existing trainers and educators in genomics and bioinformatics is insufficient with which to meet the shortfall in the Health Service workforce. There is thus an urgent need for 'Training the trainer' initiatives to enable both the academic and NHS sectors to be responsive and fit for future purpose.

There remains a need to fully explore the broader workforce demands inclusive of all potential delivery partners including industry and to dynamically model the workforce requirements and the synergistic opportunities for both training and research.

Recommendations:

- A multi-agency strategy should be developed to implement programmes that will increase training in biomedical genomics based on robust workforce modelling to provide evidenced-based, comprehensive information on and analysis of future demands.

- Re-train and develop 20% of the current 150 clinical geneticist population as practitioners of a new specialism, clinical genomics.

- Develop a new strand of medical speciality training to increase the number of clinical genomicists.

- Embed the healthcare science workforce's clinical bioinformatics speciality and career pathway including a new Masters level pre-registration training programme and a higher specialist scientific training programme .

- Continue to recruit and develop clinical scientists in genetics in biomedical genomics at all levels of the career pathway to expand and continue their work alongside clinical geneticists and genomicists

- Develop the necessary Continuing Professional Development (CPD) programmes to support the education and training of medical, nursing, science, pharmacy and managerial staff in this domain (no estimates of required numbers are yet available).

- As an urgent priority, increase postgraduate training capacity and capability in order to 'train the trainers'.

- Promote cross-fertilisation between academic, health, research and industry sectors through collaborative working on genome bioinformatics, e-health record research and variant interpretation and pursue joint training and research opportunities in collaboration with Health eResearch Centre initiatives

[i] Building on our inheritance. Genomic technology in healthcare. A report by the Human Genomics Strategy Group. January 2012. http://www.dh.gov.uk/health/2012/01/genomics/

[ii] The House of Lords Science and Technology Committee Report on Genomic Medicine (2009) http://www.publications.parliament.uk/pa/ld200809/ldselect/ldsctech/107/107i.pdf

[iii] http://www.gmc-uk.org/8___GMC_response_to_the_Shape_of_Training_Review_Call_for_Ideas_and_Evidence.pdf_51057309.pdf GMC response to the Shape of Training Review

Annex 3a

Developing the bioinformatics workforce

1. A bioinformatics workforce model was built based on the design principles set out in the Working Paper and Modelling Procedure shown in Annexes 1 and 2.

2. Preliminary analysis based on data from accredited sources, where it was available, and assumptions, based on guidance from authoritative experts, shows, under certain conditions, there will be a shortfall in the bioinformatics workforce over the next five years.

3. Assuming that all cost savings in genome sequencing occur as a result of a reduction in capital and consumables, rather than labour, a reduction in the cost of genome sequencing will result in an increase in demand without a compensating reduction in the labour needed per test.

4. It is also conceivable that the complexity of testing and analysis will increase as demand emerges for more sophisticated tests such as whole genome sequencing, currently used for research purposes only. This will increase demand for the bioinformatics work force in a way that is not reflected in the aggregate quantum of tests recorded.

5. To parameterise this scenario, it was assumed that the demand for workforce in regional centres depends purely on the trend in the number of cancer and rare disease samples tested, and not on the cost per genome sequenced. Demand for workforce in infectious disease surveillance centres was assumed to follow the same trend.

6. The preliminary analysis set out in the table below shows the supply and demand for the bioinformatics workforce between 2012/13 and 2019/20. The model suggests that all roles will face an increasing shortfall in personnel over the next 10 years:

Job description		2012/13	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19	2019/20
Clinical geneticist	Supply	91	87	83	79	79	80	80	81
	Demand	91	100	106	113	120	126	131	137
	Surplus/Shortfall	**0**	**-13**	**-24**	**-35**	**-41**	**-46**	**-51**	**-56**
Clinical genomicist	Supply	23	22	21	20	20	20	20	20
	Demand	23	25	27	28	30	31	33	34
	Surplus/Shortfall	**0**	**-3**	**-6**	**-9**	**-10**	**-12**	**-13**	**-14**
Geneticist	Supply	568	520	476	445	468	503	539	576
	Demand	568	620	663	705	746	784	819	852
	Surplus/Shortfall	**0**	**-100**	**-188**	**-260**	**-278**	**-281**	**-280**	**-276**
Bioinformatician	Supply	26	24	22	20	21	23	24	26
	Demand	26	28	30	32	34	36	37	39
	Surplus/Shortfall	**0**	**-5**	**-9**	**-12**	**-13**	**-13**	**-13**	**-13**
Public health specialist	Supply	40	37	35	32	32	32	31	31
	Demand	40	43	46	49	52	55	57	59
	Surplus/Shortfall	**0**	**-6**	**-11**	**-17**	**-20**	**-23**	**-26**	**-28**
Microbiologist	Supply	40	36	33	31	33	35	38	40
	Demand	40	43	46	49	52	55	57	59
	Surplus/Shortfall	**0**	**-7**	**-13**	**-18**	**-19**	**-20**	**-20**	**-19**

7. The analysis was constrained by time and therefore needs further development in the next revision cycle.

[i] Building on our inheritance. Genomic technology in healthcare. A report by the Human Genomics Strategy Group. January 2012. http://www.dh.gov.uk/health/2012/01/genomics/

[ii] The House of Lords Science and Technology Committee Report on Genomic Medicine (2009) http://www.publications.parliament.uk/pa/ld200809/ldselect/ldsctech/107/107i.pdf

[iii] http://www.gmc-uk.org/8___GMC_response_to_the_Shape_of_Training_Review_Call_for_Ideas_and_Evidence.pdf_51057309.pdf
GMC response to the Shape of Training Review