UCL DEAN OF MEDICAL SCIENCES
PROFESSOR DAVID LOMAS

20 March 2013

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

Dear Sally,

Many thanks for asking me to chair the strategic priorities working group for the 100k genome project. The working group considered priorities within cancer, rare diseases and infectious disease. Our selections are based on areas where we believe the introduction of genomic technology will have the greatest benefit for patient health. We recommend the following priorities for whole genome sequencing:

Cancer
Cancers will benefit from intensive genomic analyses but as they are heterogeneous a single sample from one patient's tumour will not adequately describe the biology of this disease. We believe that to make the most of NHS based genome sequencing, the patients' own genome should be included, and at least one or two additional samples from patients' cancers, separated in either space or overtime during the course of the disease. Whilst all cancer sites would benefit from genome sequencing, areas of prime importance are:

(i) Lung Cancer. This will require intensive infrastructure planning to ensure sufficient diagnostic material is available from most cases, but is likely to lead to new therapeutic opportunities and improved outcomes in the next few years

(ii) Paediatric cancers (both solid tumours and haematological malignancy). With around 1600 cases a year, and increasing long-term survivors, systematic sequencing of the UK paediatric cancer population will likely identify many new targets as well as the potential to better understand the long-term serious treatment-induced complications that, as survival continues to improve, are becoming a significant health care issue.

The infrastructure required to deliver the above is applicable to many other common cancers, such as breast, GI, prostate etc. However in addition, it offers the opportunity to increase our understanding of less frequent clinical scenarios, for example:

(i) Rare cancer syndromes. Familial clusters of uncommon cancers with no documented heritable gene defect. There is no systematic, national approach to these cases – the 100k genome project offers an excellent opportunity to collate the clinical and genomic data on these clusters to better understand causality and therapy.

(ii) Cancers of unknown primary. It is relatively common for a patient to present with metastatic disease of unknown primary origin. Conventional prognosis and therapy is largely driven by the underlying biology inherent in the site of origin, but a genomic approach to ascertaining the biology may well be quicker, easier and cheaper than a complex diagnostic work-up.

Please see Appendix 1 for further details.

Rare diseases

The implementation of whole genome sequencing within the NHS presents an exciting opportunity to deliver more comprehensive genetic diagnostic testing for rare diseases which affect ~6% of the population, many of which remain undiagnosed or imprecisely diagnosed with current diagnostic methods. This will improve the quality of NHS diagnostic services and may reduce costs by avoiding step-wise testing. Initially, whole genome sequencing should be applied in conjunction with current NHS diagnostic approaches until the specificity and sensitivity of the methods have been established. However it is envisaged that the findings from whole genome sequencing will drive new approaches to the routine use of next generation sequencing technologies. Patients should be carefully selected for whole genome sequencing based on the likelihood of making a genetic diagnosis and the associated clinical benefits. Consent for the analysis of specific regions of the genome, with reporting initially restricted to variants relevant to the patient's clinical phenotype, will partly address some of the ethical and technical challenges. Data sharing of novel variants identified within the whole genome data, linked to phenotypic information, will enhance the interpretation of current and new NHS diagnostic tests. The high diagnostic yield for rare diseases will establish proof-of-principle for an NHS-wide genomic data platform and yield immediate clinical benefits for many patients.

Please see Appendix 2 for further details.

Infectious disease

Human immunodeficiency virus, hepatitis C virus and tuberculosis have been prioritised for whole genome sequencing on the grounds of importance to human health, the strong clinical need for microbial sequencing, the availability of existing knowledge and/or infrastructure relating to microbial sequencing, and feasibility of implementation within a short timeframe. Whole genome sequencing of these pathogens will detect genes associated with resistance to antimicrobial drugs, and this information can be used to stratify therapy for individual patients. Effective therapy of individuals would be predicted to reduce the risk of spread of resistant strains. Additional public health benefit will arise from analyses of sequence data for the purposes of local and national surveillance and outbreak investigation. This will establish an infrastructure that can be applied to a much wider range of pathogens in the longer term.

Please see Appendix 3 for further details

Finally I would like to thank all members of the 'Strategic Priorities' group for all their hard work on this project. They are listed in Appendix 4 and have offered to help with implementing the proposals if that would be of assistance.

Very best wishes

Prof David Lomas PhD ScD FHEA FRCP FMedSci
Chair, Strategic Priorities Working Group

Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000

Appendix 1. Strategic Priorities in Cancer

Executive summary

Cancers will benefit from intensive genomic analyses but as they are heterogeneous a single sample from one patient's tumour will not adequately describe the biology of this disease. We believe that to make the most of NHS based genome sequencing, the patients' own genome should be included, and at least one or two additional samples from patients' cancers, separated in either space or overtime during the course of the disease. Whilst all cancer sites would benefit from genome sequencing, areas of prime importance are:

Lung Cancer

This will require intensive infrastructure planning to ensure sufficient diagnostic material is available from most cases, but is likely to lead to new therapeutic opportunities and improved outcomes in the next few years

Paediatric cancers (solid and haematological)

With around 1600 cases a year, and increasing long-term survivors, systematic sequencing of the UK paediatric cancer population will likely identify many new targets as well as the potential to better understand the long-term serious treatment-induced complications that, as survival continues to improve, are becoming a significant health care issue.

The infrastructure required to deliver the above is applicable to many other common cancers, such as breast, GI, prostate etc. However in addition, it offers the opportunity to increase our understanding of less frequent clinical scenarios, for example:

Rare cancer syndromes

Familial clusters of uncommon cancers with no documented heritable gene defect. There is no systematic, national approach to these cases – the 100 000 genome project offers an excellent opportunity to collate the clinical and genomic data on these clusters to better understand causality and therapy.

Cancers of unknown primary

It is relatively common for a patient to present with metastatic disease of unknown primary origin. Conventional prognosis and therapy is largely driven by the underlying biology inherent in the site of origin, but a genomic approach to ascertaining the biology may well be quicker, easier and cheaper than a complex diagnostic work-up.

Introduction

From the perspective of cancer, the group welcomes this initiative, committing resources to explore and understand the genomic profiles of human cancer in the NHS.

Cancer is *par excellence* an example a disease of disordered genomes, and consequently ideal for a major investment in genome sequencing. However, since human cancers are rarely the consequence of a single mutation, a whole genome sequence of a single sample from each patient without linkage to other assays will be of limited benefit in the NHS over the next 5 years. The group therefore felt quite strongly that in addition to essential linkage to clinical data, consideration needs to be given to linking genomic analyses with protein level and/or function (whether by immunohistochemistry or other assays), and with other genome analyses in the same patient across time and space to explore the heterogeneity and genomic evolution within individual cancer cases. Ideally we would have at least three samples analysed per patient's cancer. We believe that such a broad approach will be necessary to resolve the issues around drug resistance, whilst at the same time offering pivotal data that will make the UK an attractive environment in which to test new anti-cancer therapies. It can be argued that although the costs of whole genome sequencing are falling, there remains the potential that sequencing a panel of a maximum of ~500 key cancers genes, rather than whole exome or genome, would deliver more immediately "actionable" genomic data for a greater number of patients for the same cost. However, the more fundamental understanding will be delivered by whole genome sequencing, and as the cost differential between targeted gene and whole genome sequencing falls, and the bioinformatics capability expands, the whole genome approach will offer unique insights into the full landscape of cancer genomics that may not be available through other programmes.

A further benefit of whole genome sequencing of cancers is that it will allow the derivation of a unique signature for each patient of chromosomal rearrangements that can be monitored non-invasively in blood samples over time. Whilst still a research area, there is a growing interest in applying genome sequencing to plasma DNA: searching in plasma DNA for chromosomal rearrangements known to be in the primary tumour could facilitate early detection of recurrent or progressive disease.

Mindful of the requirement that this investment in genome sequencing needs to lead to improved clinical outcomes over the next ~5 years, the group felt that there were two leading candidate areas for whole genome sequencing in clinical samples, a number of other key disease areas for subsequent development, and a couple of other small but important areas that could opportunistically benefit from the infrastructure required.

Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000

Proposal for whole genome sequencing in childhood cancer (including haematological)

Cancer is the number one cause of death from disease in children in the UK, with 1,600 new cases of childhood cancer, and 260 children under the age of 15 dying of the disease each year. This means that one in 600-700 young adults is now a childhood cancer survivor, with over 33,000 childhood cancer survivors in the UK.

Long-term consequences of childhood cancer and its treatment are significant. Life expectancy is dramatically reduced in survivors with an 11-fold higher mortality rate in UK childhood cancer survivors compared to controls (Reulen *et al* JAMA **304**:172, 2010); second cancers and cardiovascular events being the major causes of premature death. In addition, cancer survivors can also suffer significant long term chronic health and developmental problems, with significant personal, social and NHS burdens. Lifelong collection of follow-up data can be done via the NHS.

The future challenge in the management of childhood cancer is to maintain and continue to improve cure rates whilst at the same time reducing or eliminating treatment-induced mortality and morbidity. A stratified medicine approach is required both for optimal targeting of the cancer, and to understand individual patient risks for toxicity. Due to the short time frame over which childhood cancers develop, intra-patient tumour heterogeneity may be lower, and the number of driver lesions fewer, than in adult malignancies. Thus it is more likely that "actionable/driver" mutations will be identified that can be targeted with novel agents, benefitting from the Paediatric Investigational Plans (PIPs) within the EU drug regulatory framework, which will increase the opportunities for inward investment by the pharmaceutical industry as well as earlier access to novel agents for UK paediatric oncology patients.

In addition to the whole genome sequencing of tumour DNA from children with cancer, there is also a compelling case for sequencing host DNA as well. There is substantial inter-patient variability in treatment-induced mortality and morbidity, a component of which will be genetic. If it were possible to predict those individuals most at risk from acute and late normal tissue toxicity including secondary malignancies, it would be possible to modify treatment and follow-up protocols accordingly to improve outcomes with less resources. A second reason for undertaking whole genome studies on host DNA in children with cancer is that it may lead to the identification of hitherto unknown very rare or medium penetrance cancer predisposition genes.

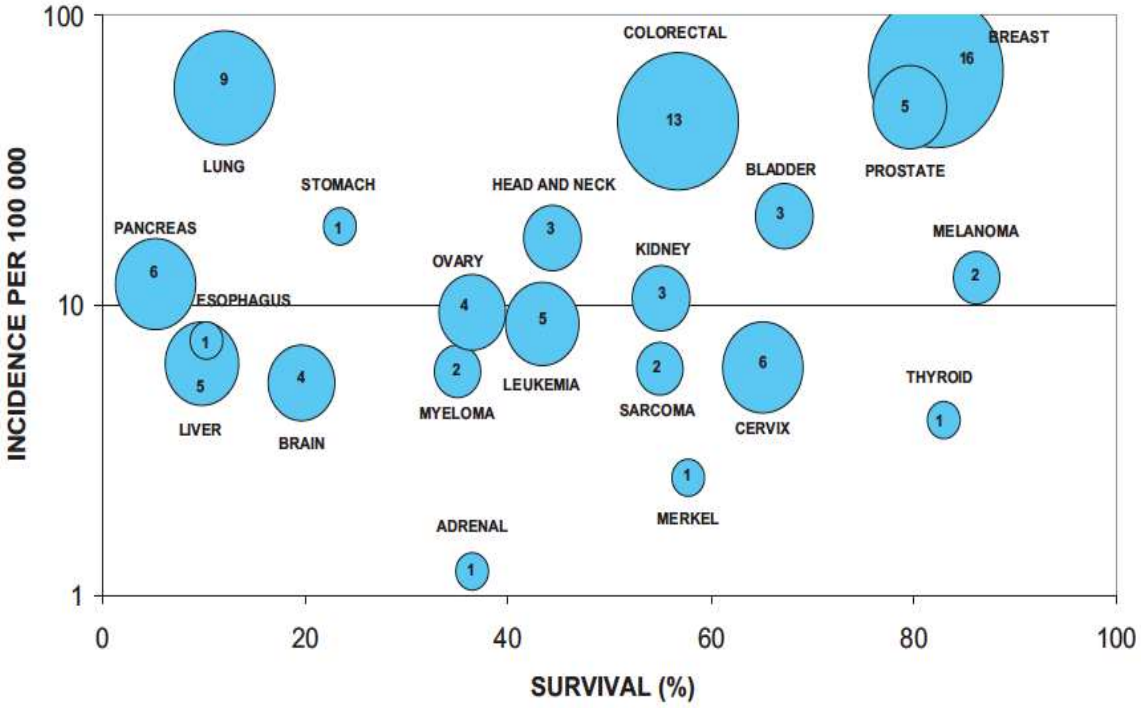
The paediatric oncology community has for many years operated an effective network of around 20 treatment centres, with long-established collection procedures for tumour samples. Thus with only 1,600 new patients each year, it will be feasible to access and sequence both the tumour and host genomes for the vast majority of UK childhood cancer cases. Access to historical collections and childhood cancer survivors would also be relatively easy and would place the UK at the forefront of genomic information in paediatric cancers, and potential host genomic risks for secondary malignancies. Insights into host risk factors for both developing cancers, and toxicity from therapy, could then potentially be explored in a focussed manner in patients with adult solid tumours.

Lung cancer

All solid cancers would benefit from the approach of systematic genome sequencing, but perhaps the disease with the greatest unmet need in this area is lung cancer. With just over 40 000 cases diagnosed each year in the UK, and 35,000 deaths per year, it is the single largest cause of cancer death. Treatments are only modestly effective, and although targeted treatments have recently become available, they are only applicable to 10-15% of the lung cancer population. As with paediatric cancers, genome analysis of both the host and the cancer would be informative, and exploration of the heterogeneity within cases would be important to direct therapy appropriately. In the short term, focussed sequencing of key (up to 500) cancer genes rather than whole exome/genome would, for the same cost, allow more patients to be analysed and thus be available for novel therapies in this devastating disease. However (*vide infra*) it is anticipated that the CRUK stratified Medicine programme will focus on this area, so we see the opportunity to complement the targeted sequencing with whole genome sequencing. Finally, lung cancers are often very heterogenous, so we would strongly advocate at least two samples being analysed at the initial presentation as well as any recurrent or relapsed disease that appears later.

Only a minority of lung cancer patients undergo resection, but such patients will be a mainstay of the target population as access to sufficient tissue is easier, and sadly many of these patients still relapse. In addition, we would want to target those patients with non-resectable disease, but at the current time, not every lung cancer patient has a tissue diagnosis, and amongst those who do, there is usually very limited material available for additional analyses. A national campaign would be needed to support this focus on lung cancer to ensure clinicians make greater efforts to obtain tissue diagnoses with sufficient material to allow genomic analyses to be done: this would be an important change in practice needed to underpin the programme.

It is important that the whole genome approach is developed as complementary to the CRUK stratified medicine programme. As the CRUK stratified medicine programme is developed, there will be a shift towards greater tissue access which is necessary for any degree of gene sequencing. Whole genome sequencing of germline and sequential tumour samples will offer insights into tumour evolution which will give rise to new therapeutic opportunities. These same genome sequencing technologies will also enable targeted gene screens of cancer



oid clinical
hus at the
will be the
of a much
nts.

the group
le system"
and lung
ta already
ady being
y building
It is likely
ess, hence
cing. This
ave similar

Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000

Carcinoma of unknown primary

This real clinical conundrum has of recent years been subjected to research in an attempt to better understand the tissues of origin. The scenario is that of a patient who presents with what is clinically and pathologically metastatic cancer, but there is no evidence of an origin in a primary tissue. Increasingly Oncology centres are defining protocols to standardise the diagnostic work-up, but despite this many cases remain without a clear primary origin. Given that the therapy of metastatic cancer depends largely on the tissue of origin, plus other molecular characteristics, a panel approach to genomics which allowed rapid identification of the likely tissue of origin would influence therapy, and in some cases, identify patients whose disease can benefit substantially from a particular therapeutic approach which would not otherwise be evident from standard diagnostic approaches. There are no clear incidence data, since the frequency of this presentation depends on the intensity of the diagnostic work-up: a single genomics approach would potentially be quicker, easier and potentially cheaper than the series of different imaging and serological tests that underpins the current patient management.

New familial cancers/rare presentations

From time to time individual centres see rare cases - but there is no system to collate these cases. The infrastructure required to deliver the above, particularly for lung cancer, could be applied to any cancer patient, and if made available for new rare cancers, or clusters of cancers amongst relatives without a recognised familial risk, would allow the UK to create a virtual resource for new rare cancer cases with complete tumour and host genome sequencing and longitudinal clinical data. Given the benefits of data collation in a single health care system with national coverage, this would place the UK at the forefront of such genetic research.

We would propose therefore a central data collection repository, managed by the NHS, into which clinicians with patients' consent could report unusual familial collections of cancers. Samples would be made available for genome sequencing, and this virtual collection could be overseen by an expert group involving both geneticists and oncologists which would decide when a familial cluster was sufficiently unusual to subject the patients' own and tumour samples for analysis. It is difficult to define the criteria in advance, but for example a cluster of 3 or more less common cancers in one family would trigger an analysis, and for rare cancers two might be sufficient.

Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000

Appendix 2. Strategic Priorities in Rare Diseases

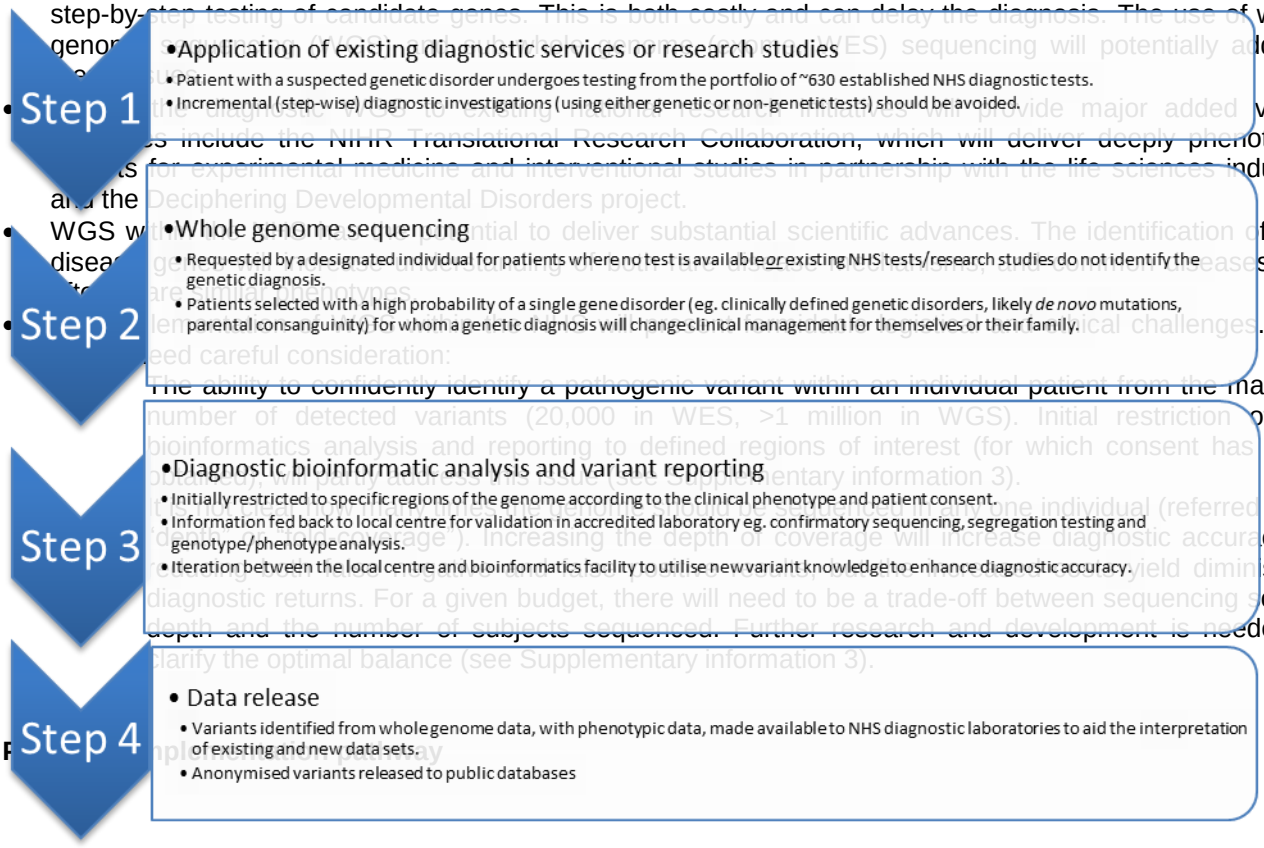
Group members: Patrick Chinnery, David Lomas, Helen Firth, John Bradley, Mark McCarthy, Sian Ellard, Dan Bridge, Cathleen Schulte, Mark Bales

Executive summary

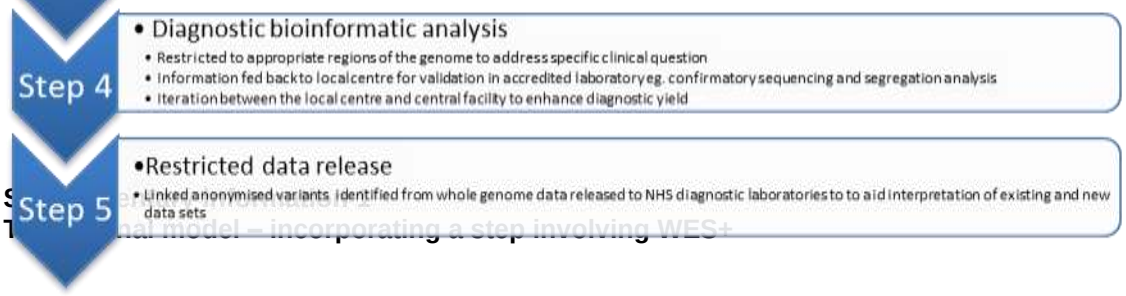
The implementation of whole genome sequencing within the NHS presents an exciting opportunity to deliver more comprehensive genetic diagnostic testing for rare diseases which affect ~6% of the population, many of which remain undiagnosed or imprecisely diagnosed with current diagnostic methods. This will improve the quality of NHS diagnostic services and may reduce costs by avoiding step-wise testing. Initially, whole genome sequencing should be applied in conjunction with current NHS diagnostic approaches until the specificity and sensitivity of the methods have been established. Patients should be carefully selected for whole genome sequencing based on the likelihood of making a genetic diagnosis and the associated clinical benefits. Consent for the analysis of specific regions of the genome, with reporting initially restricted to variants relevant to the patient's clinical phenotype, will partly address some of the ethical and technical challenges. Data sharing of novel variants identified within the whole genome data, linked to phenotypic information, will enhance the interpretation of current and new NHS diagnostic tests. The high diagnostic yield for rare diseases will establish proof-of-principle for an NHS-wide genomic data platform and yield immediate clinical benefits for many patients.

Recommendations

- Rare diseases present an ideal opportunity to establish a platform for the application of high-throughput genomics in routine NHS practice. As a group, rare disease affect 6% of the UK population, and >85% are caused by a single gene defect. Many are chronic, and associated with substantial morbidity and premature mortality. Early diagnosis enables accurate genetic counseling and prevention, and may lead to new treatments based on genetic stratification. Inherited cancer and immunodeficiencies fall within this group.
- There has been an exponential growth in genetic knowledge within the last three years (~27 new diseases per month in OMIM in 2012 for which the gene has been identified). This presents a major challenge to NHS diagnostic services. At present, testing for >600 different disorders is available within the NHS diagnostic laboratory network (UKGTN) but this represents <1/4 of known disease genes. Current practice is based on step-by-step testing of candidate genes. This is both costly and can delay the diagnosis. The use of whole genom… Deciphering Developmental Disorders project.
- WGS w… disea… s that …
- … Two …



Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000



Transitional model – incorporating a step involving WES+

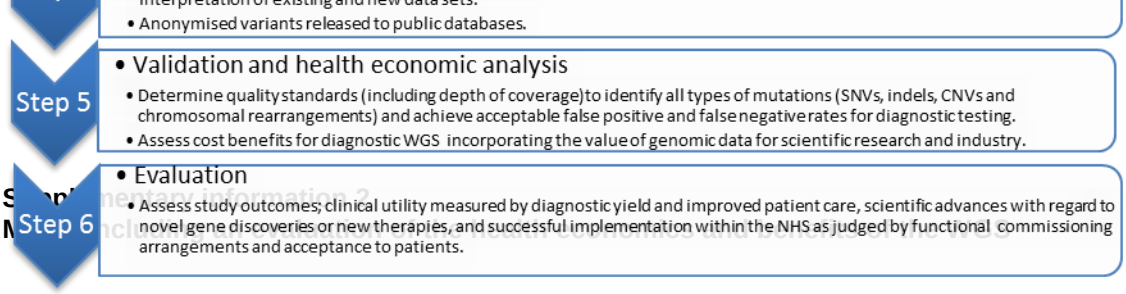
A transitional model would enable the 100k genomes project to be established rapidly (within the projected time course), and would build on the pilot work of the DDD project to ensure that:

(i) Results are provided for patients in a timely fashion (eg. within 8 weeks) and with sufficient clinical accuracy (not yet established for WGS)

(ii) Research output (eg. new gene discovery and drug targets) is maximised in the short-medium term (based on higher volumes of interpretable data generating greater power by maximising the chance of seeing the same functional element hit twice or more in patients with similar clinical phenotypes).

(iii) The research and development work necessary to ensure adequate clinical performance of WGS is undertaken, and the infrastructure necessary to deliver this innovation is established. Once validation experiments show equivalent or superior diagnostic accuracy of WGS, step 2 (WES+) could be omitted.

Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000



Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000

Supplementary information 3
WGS and clinical value for rare disease diagnostics.

One of the key issues which has emerged in discussions within the rare diseases group centres on the clinical value of WGS for rare disease (RD) mutation screening. The specific focus has been on understanding how the performance of WGS data compares with that of diagnostic sequence data currently generated within the NHS. Those clinical sequence data have historically been generated with Sanger sequencing and MLPA, though there is increasing use of next-generation sequencing methods to interrogate panels of target genes.

Rationale
WGS data has obvious potential to provide clinical benefits through novel discoveries, including the detection of causal mutations for RDs that have, as yet, no known cause, and the identification of novel transcripts harbouring causal mutations for RDs where existing knowledge is incomplete.

However, to maximise the clinical value of the WGS 100k program, it would also be highly desirable for the WGS data to be of sufficient quality to substitute for existing sequence-based diagnostic protocols; or, where relevant tests/panels do not yet exist for a given indication, to generate data that would allow WGS sequencing to provide an accelerated route for expanding the diagnostic portfolio.

In one model proposed at the first CMO-GAG meeting, Helen Firth had suggested that it might be desirable, at least initially, to parse WGS data from RD patients into two components. The first part would be made up of the established set of target genes for that particular clinical indication: that is, those genes, candidates for inclusion in an existing or putative targeted sequencing panel, for which there is already sufficient prior evidence to support clinical decision making. The remainder of the WGS sequence would, in most instances, be of less immediate clinical value for any given patient, but would offer a valuable resource for further research, including the identification of additional causal mutations involving novel genes in those in whom the target panel screen is negative.

Such a division into “clinical” and “research” parts would have obvious advantages. For example, it might allow rapid integration of WGS data into clinical genetics service delivery by focusing on the existing knowledge base for each indication: it would be less contingent on the development of the informatic systems, ethical protocols and medical knowledge required for clinical interpretation of the full genome. Another advantage would be the “one size fits all” argument: there would, in principle, be no need to go to the trouble of developing and validating novel “capture” assays for each indication-specific target panel if WGS could deliver sufficient coverage for any given set of genes of interest.

What kind of WGS would be required?
However, for this arrangement to work, the “target” sequence filleted out from WGS needs to be of sufficient quality to meet the exacting demands of clinical sequence diagnostics. Stark binary outcome decisions (termination, prophylactic surgery) often hinge on the interpretation of the sequence findings, leaving little tolerance for error. Formal adoption of any such test into NHS practise involves detailed evaluation[1] to ensure that clinical decisions are based on allele calls that have:

- very low false negative rates, to provide confidence that all bases of interest in the target gene have been interrogated at sufficient depth, and variants well-enough called, to ensure reliable detection of heterozygote alleles of all classes, and equally importantly, to provide confident exclusion of the target when no variant alleles have been detected: whereas a sensitivity of 95% is often deemed acceptable for research studies, diagnostic tests aim for >99% sensitivity; and
- low false positive rates, to keep the effort required to validate all putative causal variants under some control: otherwise the validation effort can rapidly become onerous and/or false assertions of causation made. Note that the false positive rate is highest for those rare alleles which appear to have the most severe effects on gene function[2] i.e. those with the strongest credentials as causal.

It is also worth remembering that to be clinically useful, a definitive result needs to be provided to the clinician and patient within an acceptable period (usually no more than a few weeks).

In the case of NGS-based targeted sequencing, very high depths of sequence coverage are possible at relatively low additional cost (since the sequencing costs are typically only a fraction of the costs of the capture). This means that very deep coverage is readily attainable at the vast majority of ascertained sites, even allowing for the additional unevenness of coverage associated with any capture method (though this can to some extent be mitigated by adjusting bait density/tiling or by using alternative PCR-based targeting methods).

Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000

In the case of WGS data, coverage is generally more even, but the costs of boosting average read depth are substantial.

What we would really like to know is: what level of average WGS coverage would be required, such that for the vast majority of clinical target sets, acceptable FN and FP rates can be achieved?

What depth of WGS would be required to "replace" current diagnostic methods?
There is a widespread view that a 30x genome is sufficient for diagnostic purposes, but as with much "conventional wisdom", the basis for this is not entirely clear. Where the aim is research, discovery WGS to a mean depth of 30x will often be ample, but most NGS assays for diagnosis target at least 30 reads for each base to achieve an equivalent sensitivity to the Sanger-based tests they replace. Typically, where no pathogenic mutation is detected, regions with coverage below 30x are subjected to Sanger sequencing to ensure no mutation has been missed. Having said that, the CMGS guidelines[1] do accept that the required sensitivity of the test will depend on the clinical application and a lower sensitivity may be acceptable for a large panel (>20 genes) in which the prior likelihood of finding a mutation is lower, and testing has not been previously available. Some target panels currently being used accept lower minimum base coverage (>18-20x). Note also that there is an important distinction between minimum coverage per base and mean coverage. For example, in the Exeter capture based NGS gene panel a mean coverage of ~180x was required to achieve a minimum 30x coverage.

The question of "adequate" WGS depth turns out to be rather difficult to answer for a number of reasons. First, there is relatively little very high depth whole genome sequence data out there that has been used to look at this question. Second, an understanding of FP rates requires extensive validation studies that have not yet been widely performed. Third, the coverage required will depend on the type of variant involved: indel calls from NGS remain far harder than SNP calls[3], and liable to much higher FN and FP rates. (Note that since ~40% of protein truncating variation is due to fs indels[2], for clinical purposes, accurate indel calling is essential). Fourth, alternative WGS platforms have different profiles in terms of the variants detected[3]. Finally, there will be instances where WGS has specific advantages over targeted sequencing[4]: most obviously, where causal alleles outside the target region, or where target capture is incomplete, but also in the case of larger insertion/deletions or translocation/inversions.

For these reasons, most studies that have looked at this question have focused on the relationship of WGS depth to the detection of known SNV alleles (i.e. sensitivity and FN rates). From this limited perspective, there are some indications that a 30x WGS genome might not be too wide of the mark, but that 40-50x average coverage would likely be safer.

Gil McVean has generated some estimates based on the assumptions that (a) the base of interest lies in the accessible genome (not always true of course, but probable for a variant already implicated in disease risk); (b) three reads of the variant allele is generally sufficient for accurate calling of SNVs (though not indels) and that allele balance is binomial; and (c) WGS data read-depth distribution is Poisson-like but with twice the variance (the last of these based on empirical observation in WGS data generated in Oxford). On those grounds, 30x coverage should guarantee that at least 3 variant reads at ~98% of the genome (and hence an average of 98% of any chosen target) ie at 99% of sites there is a 99% chance of seeing at least 3 variant reads if the individual is heterozygous. Note that this would not, even for SNVs, reach the >99% sensitivity typically sought for diagnostic tests.

Empirical data analyses suggest that these estimates may be on the optimistic side, even for SNVs, though this may in part reflect the fact that the empirical data are rather old, and sequencing and calling protocols have improved in the interim. For example, Ajay and colleagues[5] report that "an average mapped depth of 50x…was required..to produce confident [SNV] genotype calls for >94% of the genome and >80% of the coding exome" on GAII, but that on the more recent HiSeq2000 machines and updated sequencing protocols, the figures [SNV only] for both genome and exome reached ~95% once the average read depth approached 30-40x (ibid, Figure 6).

Lam and colleagues at Stanford[3] compared ~75x GWS data generated on both the Illumina and CGI platforms, and found a high proportion of platform-specific variants (only 88% of all SNV calls and 26% of indel calls were concordant across the two). Whilst most concordant SNV calls were validated with orthogonal technologies, platform specific SNV calls were almost evenly distributed as FN and FP, and the figures were worse for indels. These data indicate substantial FN and FP rates persist (even for SNVs) for any single sequencing platform, despite 75x coverage.

Taken together these data would tend to suggest that WGS coverage well in excess of 30x is required to constrain FP and FN rates to acceptable levels, especially for indels. However, further work is required to

interrogate available high-depth WGS and target sequence data (particularly data generated more recently) to understand more completely the relationships between WGS coverage depth and FN and FP rates for different variant types.

Interim conclusions

From the perspective of the RD subgroup of the scientific priorities, the following statements seem reasonable at this stage:

- to be of clinical diagnostic value, WGS data needs to be able to detect (and exclude) clinically actionable variants of all classes (SNVs, indels, translocations, large deletions etc) with very high sensitivity and specificity;
- pending more detailed empirical data, it seems likely that average WGS depth well in excess of 30x (possibly as high as 100x) would be required to provide coverage of diagnostic targets sufficient to ensure acceptable FN and FP rates, at least with current technologies and calling algorithms;
- failure to constrain FP rates (as well as FN rates) would lead to an explosion in the effort that has to be expended in the labor-intensive validation of putative causal variants: the associated validation costs have the potential to swallow up any apparent savings made by reductions in WGS sequencing depth;
- if high-depth sequencing is not feasible, and if technological and computational improvements are not forthcoming, then WGS data alone will not be able to replace existing gene-based and panel-based tests for most, if not all, established clinical indications: for the foreseeable future, these and other mutation detection assays would have to be performed in parallel, so as to deliver acceptable clinical performance.

Mark McCarthy
Feb 2013

In preparing this summary, I have consulted with the following:
· Jenny Taylor, Oxford
· Gil McVean, Oxford
· Euan Ashley, Stanford
· Nazneen Rahman, ICR
· Sian Ellard, Exeter

References

1. Ellard S et al (2012) Practice Guidelines for Targeted Next Generation Sequencing Analysis and Interpretation (available at http://www.cmgs.org/BPGs/BPG%20for%20targeted%20next%20generation%20sequencing%20final.pdf)
2. MacArthur D et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335;823-8;
3. Lam H et al (2012). Performance comparison of whole-genome sequencing platforms. Nat Biotech 30;78- 82;
4. Clark MJ et al (2011). Performance comparison of exome DNA sequencing technologies. Nat Biotech 29;908-914;
5. Ajay S et al (2011). Accurate and comprehensive sequencing of personal genomes. Genome Research 21;1498-1505.

Appendix 3. Strategic Priorities in Infectious Diseases

Members:
Sharon Peacock (chair)
David Livermore
Julian Parkhill
Deenan Pillay
Grace Smith
Brian Spratt

Proposed scientific priorities for infectious diseases:
1. Human immunodeficiency virus (HIV)
2. Hepatitis C virus (HCV)
3. Tuberculosis (TB)

Executive summary:
Human immunodeficiency virus, hepatitis C virus and tuberculosis have been prioritised on the grounds of importance to human health, the strong clinical need for microbial sequencing, the availability of existing knowledge and/or infrastructure relating to microbial sequencing, and feasibility of implementation within a short timeframe. Whole genome sequencing of these pathogens will detect genes associated with resistance to antimicrobial drugs, and this information can be used to stratify therapy for individual patients. Effective therapy of individuals would be predicted to reduce the risk of spread of resistant strains. Additional public health benefit will arise from analyses of sequence data for the purposes of local and national surveillance and outbreak investigation.

Overview of process for selection of priorities:
The expert group taken as a whole contained broad expertise, including knowledge of a wide spectrum of infectious diseases that occur in the community and hospitals, in the UK and globally. This included an extensive knowledge relating to infections caused by bacteria and viruses, and the growing problem of drug resistance. The group included clinicians involved in patient care, academic clinicians, and non-clinicians with direct and recent experience of the application of whole genome sequencing to microbes. The group also contained expertise in bioinformatics analyses, genome sequence interpretation software, and web-based interfaces for interpretation of sequence data and collation of epidemiological and clinical information. Two members (BS and JP) were involved in a recent external review of the bioinformatics needs of the Health Protection Agency (BS was committee chair). This heard evidence from many academic and other groups, and published a 27-page document recording the sequencing capacity and translational work being done at the HPA and elsewhere - including representation from the three Wellcome Trust/Department of Health funded HICF (Health Innovation Challenge Fund) projects on whole pathogen sequencing of bacteria and viruses. This report was a valuable source of information for the CMO Science group.

In response to guidance from the Department of Health, we aimed to select around three priority organisms that would provide proof of principle for genome sequencing technology, could be implemented immediately or in the near future, would bring clear clinical benefits, and would promote the establishment of an infrastructure that would allow this technology to be applied to a much wider range of pathogens in the longer term.

We considered the utility of whole genome sequencing for the management of patients with viral illnesses, and the potential for an initial small selection of pathogens to provide the stimulus and infrastructure for future expansion to a wider range of organisms. This included discussion of the human immunodeficiency virus (HIV), hepatitis C virus (HCV), hepatitis B virus, influenza and norovirus. There was unanimous agreement to focus on HIV and HCV. The basis for this decision is presented in ‘key points’ and ‘justification for inclusion’ below, but in summary was based on a combination of: clinical importance; a clear need for sequencing in the clinical care pathway; existing clinical experience of the use of sequence data (particularly for HIV); an existing framework that would support implementation; the provision of information on drug resistance that would guide therapy for the individual patient; and the provision of information of public health importance relating to viral transmission. Hepatitis B was not included because the clinical utility of whole genome sequencing is not as strong as for HIV and HCV, although this decision could be reversed if full-length HBV sequencing gained a stronger role in patient care. Influenza was not prioritized because of the sporadic nature of epidemics/pandemics, combined with the extensive existing resources available to tackle this infection. Norovirus is also sporadic, and is a self-limiting illness for which an outbreak event is often obvious based on epidemiological information. Nevertheless, we felt that the initial development of a core sequencing capacity could also deal with seasonal surge capacity for viruses such as influenza and norovirus over time, with the benefit of enhanced infection control and epidemic mapping.

Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000

We considered a range of bacterial diseases, and debated the potential short-term impact that whole genome sequencing of each species or group of species could have on patient care and public health. This included consideration of tuberculosis (to determine pathogen transmission and predict antimicrobial drug susceptibility), *Clostridium difficile* and methicillin-resistant *Staphylococcus aureus* (MRSA) (to determine transmission and strain evolution in healthcare facilities), *Salmonella* and *Shigella* spp. (to confirm an outbreak of food poisoning), *Neisseria gonorrhoeae* and multi-drug resistant Gram-negative bacilli (for rapid genetic prediction of drug susceptibility), and *Escherichia coli* (to better understand the rising number of bloodstream infections, multi-resistant or not, caused by this pathogen). There was unanimous agreement to focus on tuberculosis (TB). The basis for this decision is presented in 'key points' and 'justification for inclusion' below, but in summary was based on a combination of: clinical importance, including rising rates of TB in some groups and rising rates of drug resistance; the provision of information on drug resistance that would guide therapy for the individual patient and that could be produced much more rapidly for a range of drugs than is currently possible for this slow-growing pathogen; the provision of information of public health importance relating to transmission and outbreak investigation, which could again be produced much more accurately than is currently possible; and tractability, with a predicted number of cases per year of around 9,500. Other bacterial pathogens did not rank as highly based on this combination of features, even though individual aspects of a particular bacterium or group of bacteria were considered to be very important. For example, identification of drug resistance in *Neisseria gonorrhoeae* was considered extremely important, but the current sequencing platforms would not generate the necessary sequence data to provide a genetic prediction of drug resistance before the patient left the clinic and so would primarily benefit research, rather than clinical practice. Similarly, rapid sequencing of bacteria such as *C. difficile* and MRSA is certainly likely to be of value in investigating nosocomial transmission, whilst sequencing of *Salmonella* and *Shigella* will be important in identifying food-borne outbreaks, but the clinical and epidemiological value of these is only likely to be realised when sequencing is in widespread routine use, after this initial pilot phase, and for this reason they were de-prioritised. These assessments would change if sequencing of pathogens could be performed swiftly from the initial specimen, without the need for culture.

Sequencing capacity for infectious diseases was concentrated on the pathogen and the optimisation of antimicrobial regimens, rather than the human genome. This is because the most important determinant of recovery from many infectious diseases is whether the antimicrobial drug given is effective. In addition, public health investigations require the identification of transmission chains, which can only be achieved from pathogen genomic information. Host genome sequence is currently unlikely to provide clinically actionable data in relation to infectious diseases as, except in a few extreme cases, the host genome remains poorly understood in relation to predicting the probability of drug reactions, disease progression and outcome.

KEY POINTS

Human immunodeficiency virus (HIV)

HIV is prone to genetic mutations, some of which will be expressed in the viral proteins targeted by antiretroviral agents and result in HIV drug resistance. Once commenced on anti-retroviral therapy, patients take treatment for life and resistance testing is an integral component of the care pathway. This is performed pre-therapy, in the event of treatment failure, or when anti-retroviral therapy is modified for other reasons. Sequencing is currently performed using capillary sequencers in numerous NHS laboratories, and generates fragments of the genome. The method used is time-consuming and involves numerous steps to amplify and sequence multiple gene fragments that are then pieced together using bioinformatic tools. This could be made more efficient through the use of next-generation sequencing platforms, which would bring added value since this provides full-length sequence in a single reaction, and can also demonstrate the presence of resistance in minority virus populations. A website has already been developed through an academic initiative (UK HIV Database) which accepts and interprets sequence data and collates genetic and clinical information. Minor modifications have already been made to allow this website to accept and interpret full-length viral genome sequence data. Inclusion of HIV in our list of priorities is supported by the strong clinical need for sequencing, the potential to piggyback new sequencing initiatives onto an existing system, and represents a target with rapidly achievable goals. Full-length viral sequencing will also extend the opportunity to undertake phylogenetic analyses, which would bring a greater understanding of HIV transmission patterns.

Hepatitis C virus (HCV)

All patients with Hepatitis C infection routinely have viral genotyping at first presentation, as the genotype has implications for choice of therapy and clinical response to this. HCV genotyping during the administration of current and newly licensed directly acting antiretrovirals (DAAs) is potentially important but not part of standard care, and the development and introduction of a genetic testing protocol would represent an improvement in patient care. As with HIV, replacement of current testing platforms and protocols with newer technology would be associated with simplification of methodology, and would provide information on the presence of minority virus populations containing gene mutations that are indicative of drug resistance. Analysis of HCV genomes are drawing on existing frameworks for the analysis of other viral gene sequences (such as HIV), which could reduce the time taken to bring into clinical practice the necessary sequence analysis, generation of clinical reports, and collation with clinical data. Full-length viral sequencing will provide the opportunity to undertake phlyogenetic analyses and would be predicted to bring a greater understanding of HCV transmission patterns within and beyond England. Access to a sequencing pipeline for HCV would add value to clinical trials of DAAs conducted in the UK, as assessing the emergence of resistance is an important component of such studies.

Tuberculosis

Two-thirds of notified cases of tuberculosis are confirmed by growth of the organism from patient samples. These are initially cultured in local NHS laboratories and when positive for presumptive *Mycobacterium tuberculosis*, are referred to a reference laboratory for identification, antimicrobial susceptibility testing and epidemiological typing. It typically takes two to three weeks to culture and identify *M. tuberculosis*, and between one to two months to complete susceptibility testing and genotyping using a range of both traditional culture-based and molecular techniques. The development of molecular tests that simultaneously detect and identify *M. tuberculosis* and provide antimicrobial resistance data has improved the speed although not the sensitivity of TB diagnosis, and only target the most frequent genetic mutations for resistance to a limited number of first- and second-line antibiotics. Genome sequencing of *M. tuberculosis* has been undertaken in research settings and could be used for identification, prediction of antimicrobial drug susceptibility and epidemiological typing, but does not currently form a component of routine clinical care. Its introduction in association with the development of sequence interpretation tools could simplify workflows, reduce turnaround time for identification and genotyping, and gene-based susceptibility predictions could be based on the entire genome rather than relying on PCR or other methods that target specific regions. This should facilitate the rapid institution of appropriate therapy (which is particular importance for patients infected with multidrug resistant or extremely drug resistant strains), and would provide more accurate evidence of transmission of *M. tuberculosis* and tuberculosis outbreaks so as to inform public health interventions and monitor their efficacy.

JUSTIFICATION FOR INCLUSION OF HIV

1. *Burden of disease and existing need for sequencing:*
   (i) An estimated 100,000 individuals are infected with HIV in England and Wales, with 73,000 currently diagnosed. Around 70% of these individuals are receiving antiretroviral therapy, and total cost of care exceeds £1b per annum. In 2012, there were a record number of new infections acquired within the UK and the total burden of infection will continue to grow.
   (ii) Antiretroviral therapy (ART) for the treatment of HIV infection has improved steadily since the advent of potent combination therapy in 1996, but a major challenge to treatment is the emergence of resistance.

HIV replication is error-prone due to the absence of an effective proofreading mechanism which gives rise to a high number of mutations, some of which will emerge in the viral proteins targeted by antiretroviral agents and result in HIV drug resistance. All patients with HIV already have partial viral sequencing undertaken at the time of diagnosis or prior to antiretroviral therapy in order to exclude transmitted drug resistance, and to optimise first line treatment. Viral sequencing is subsequently undertaken at time of therapy failure. We estimate that around 15,000 such sequences are generated within the UK per annum, and this is the sample from which selection for full-length sequencing would be undertaken.

2. *Expected impact of full-length viral sequencing on patient care and public health:*
   (i) Generating higher quality data. The implementation of next-generation sequencers (for example, using a Illumina MiSeq or Ion Torrent instrument) in the care pathway of patients with HIV infection would mean that full-length genome sequences would be generated, rather than the gene fragments generated by the capillary sequencers in current use. This would provide a full prediction of resistance to the 25 available drugs (rather than for just a proportion provided by partial sequencing) before starting therapy. In addition, co-receptor tropism will allow a further estimate of the likelihood of disease progression. There is good evidence that minority viral populations that contain mutations encoding drug resistance may be present in a given individual. This is important since a minority resistant population is likely to become the majority population by selection for the variant during drug treatment. Mixed viral populations are not detected using the capilliary sequencing method in current use, but use of next-generation sequencing technologies would provide information on the presence of minority virus populations. The coverage (the number of times that the genome is sequenced in a given reaction) required to reliably achieve this is approximately 1000x, which would be predicted to detect a population present in a proportion of 1%.
   (ii) Which patients would benefit. All newly diagnosed patients with HIV infection would be expected to benefit as soon as they entered the care pathway. Ideally, all known HIV infected patients who required viral genome sequencing at the start of, or during therapy would be investigated using next-generation sequencing platforms. If the number of sequencing runs available for HIV testing was capped in the first instance, it would be possible to develop guidelines for more targeted use.
   (iii) Generating data of public health importance. HIV incidence is increasing in the UK. A number of phylogenetic methods are available to fully exploit full-length HIV sequences to describe the dynamics of viral spread, and inform public health measures for control.

3. *Opportunities for academic and commercial research:*
   (i) New genome sequence knowledge. Full-length sequencing of HIV could uncover new determinants of disease pathogenicity and drug susceptibility. This could be of utility to the pharmaceutical industry, and may benefit UK wealth.
   (ii) Sequencing the human genome. Routine clinical HIV care already includes host genome HLA B-5701 polymorphism testing in order to predict hyper-susceptibility to abacavir, one of the antiretroviral drugs. A number of other potential pharmacogenomics mechanisms are likely to be uncovered by human genome sequencing, such as the prediction of drug levels that can vary between individuals because of variability in metabolism. In addition, a number of polymorphisms - classically at the CCR5 locus - are associated with disease progression rate and even susceptibility to infection. With the current (limited) state of knowledge, however, we consider that human genome sequencing in individuals with HIV infection represents an area for research rather than clinical implementation.

4. *Linkage with existing capacity and capabilities – key considerations for data workstream:*
   (i) An existing network of laboratories that perform viral gene sequencing. Currently, 17 NHS virology laboratories across England have the capability to sequence partial length HIV genomes. Different genes are sequenced at different times, in accordance with patient therapy. These laboratories would be anticipated to be capable of running next-generation sequencing platforms. Replacement of current testing platforms and protocols with newer technology would be associated with simplification of methodology. The cost per complete test should be comparable to existing costs (excluding the capital costs of instrument purchase).
   (ii) An existing mechanism for analysis of HIV gene sequences and reporting of drug resistance to clinicians. The UK HIV Drug Resistance Database is a central repository for the results of resistance tests that are performed as part of routine clinical care in a distributed model throughout the UK. This provides the software to assemble gene fragments generated by capillary sequencers, interpret the result and generate a clinical report on the presence of resistance mutations. The modifications required to analyse full-length sequence rather than gene fragments have already been made, and the UK HIV Drug Resistance database could accept data generated from next-generation sequencers. The methodology for full-length sequencing is addressed within the ICONIC programme on full-length viral sequencing (Health Innovation Challenge Fund).

(iii) An existing mechanism for curation of sequence data and linkage to clinical data. The UK HIV Drug Resistance Database already provides linkage to a major clinical data set held at the MRC Clinical Trials Unit, and also provides the HPA with epidemiological information. Full-length sequence data would be dealt with in the same manner. Data are also linked to a clinical phenotype cohort study (Collaborative HIV Cohort Study- CHIC). Both initiatives are currently funded by an MRC Programme Grant.

(iv) Validation of the phenotypic effect of novel gene mutations. A number of research laboratories in the UK already have the capacity to undertake phenotypic HIV drug susceptibility testing in order to test whether newly found genetic mutations are associated with drug resistance.

5. *Key considerations for the ethics workstream*:

HIV infected patients have had strong advocacy groups since the onset of the epidemic, based mainly on responding to the stigma associated with the disease. This has led to a high level of engagement between HIV service users and the NHS. In general, advocacy groups have been highly supportive of research to improve the lives of infected individuals, although implementation of any new methods of testing, monitoring or intervention will be scrutinised. There are two major considerations:

(i) Identification of transmission chains. The use of HIV gene sequences to uncover the dynamics of spread through molecular epidemiological techniques has on occasion been diverted into exploring the actual donor of a newly infected individual. At its extreme, this approach has contributed to evidence brought in support of prosecutions for "reckless sexual transmission of serious infections". Despite a clear refutation of the worth of such uses of viral gene sequence data (e.g. Bernard et al. National AIDS Trust, 2007; Pillay et al. BMJ 2005), and a subsequent guidance from the Crown Prosecution Service, (http://www.cps.gov.uk/legal/h_to_k/intentional_or_reckless_sexual_transmission_of_infection_guidance/), there remains considerable anxiety about such uses of gene sequence data from the infected community. This can be mitigated by reassurance regarding the governance and use of such data.

(ii) Availability of data for new interpretation. People having full-length viral sequencing undertaken will need reassurance that these data can be interrogated should a new viral genetic predictor of, for instance, disease progression, drug toxicities, or drug resistance be identified.

Mitigation of all these potential concerns can best be addressed through active engagement with one or more of the established HIV advocacy groups (e.g. Terrence Higgins Trust, National AIDS Manual, etc).

JUSTIFICATION FOR INCLUSION OF HCV

1. *Burden of disease and existing need for sequencing:*

(i) The most recent national estimate suggests that around 216,000 people in the UK have chronic hepatitis C (160,000 people in England), an estimated 87% of whom are current or past injecting drug users. In England, more than 95,000 individuals had been diagnosed with hepatitis C by the end of 2011, suggesting that a significant number of infections remain undiagnosed. Six major genetic types of HCV have been identified, together with numerous subtypes. Genotype 1 is the most common in the UK, and is found in about 40–50% of cases. Genotype 3 contributes about another 40–50%, and genotypes 2, 4, 5 and 6 constitute the remainder of about 10%.

(ii) Early diagnosis and treatment can clear infection and reduce the risk of long-term complications. The current standard treatment for moderate and severe chronic HCV infection in adults is combination therapy with ribavirin and either peginterferon alfa-2a or peginterferon alfa-2b. Many people find this very hard to tolerate, and there are significant problems of dropout and non-adherence with treatment. Two drugs (telaprevir and boceprevir) are the first so-called direct-acting antiviral agents (DAAs) approved for use in HCV treatment. Either drug is recommended for use in combination with peginterferon alfa and ribavirin as a possible treatment for genotype 1 chronic hepatitis C in adults with the earlier stages of liver disease who are previously untreated or in whom previous treatment with interferon alfa has failed. We estimate that at least 1000 patients will be treated with the new DAAs over the next year, and that numbers treated will increase over time because of increasing diagnoses and disease progression. Treating a subset of infected patients with first generation DAAs will cost the NHS an estimated £96 million/year, with an expected treatment failure rate of 30-50%.

(iii) All patients with HCV infection require viral genotyping at the time of diagnosis as this provides an indication of the likely rate of disease progression, response to specific therapies, and identifies those patients with genotype 1 infection who could receive the currently available DAAs. The rapid replication rate of HCV together with the error-prone polymerase activity leads to a high genetic diversity among HCV genomes that includes mutants with reduced susceptibility to DAA-therapy. These resistance-associated variants often occur at very low frequencies but selection of resistance mutations may occur during DAA-based treatment, leading to treatment failure. Detection of viral mutations at the time of DAA treatment failure is required to determine the presence of possible drug resistance, and likelihood of response to future second-line therapy. Currently, viral genotyping is undertaken by a variety of techniques, including partial viral sequencing or hybridisation-based blot assays; detection of drug resistance mutations is achieved by viral genome sequencing using capillary sequencing technology.

(iv) As HCV genotype 3 does not respond to currently licensed DAAs, patients infected with this genotype continue to receive interferon alfa and ribavirin. Although response to current therapy is higher than for genotype 1, lack of information about the variability in genotype 3 genomes prevents any prediction on treatment success or failure.

2. *Expected impact of full-length viral sequencing on patient care and public health:*
   (i) Generating higher quality data. The implementation of next-generation sequencers in the care pathway of patients with HCV infection would mean that full-length genome sequences would be generated, rather than the gene fragments generated by the capillary sequencers in current use. This would provide a baseline sequence for all genes that are or could be targeted by current and future DAAs, as compared to the current approach of sequencing specific genes. As for HIV, minority viral populations may be present in a given individual and may contain mutations encoding drug resistance; this could be detected using next-generation sequencing platforms.
   (ii) Which patients would benefit. HCV sequencing during the administration of DAAs is not widely available, and there is little consistency in its use for this purpose across the country. The development and introduction of a genetic testing protocol would be a logical step if the newer technologies were to become available, and would enhance individual patient care. All newly diagnosed patients with HCV infection would be expected to benefit as soon as they enter the care pathway, when virus variability and genotype would be assessed using full-length sequencing. Additional sequencing would be used at the start of DAA therapy and in the event of treatment failure.
   (iii) Generating data of public health importance. Full-length sequencing will provide the opportunity to undertake phylogenetic analyses and would be predicted to bring a greater understanding of transmission patterns. For instance, despite recognition that on-going spread of HCV amongst HIV co-infected individuals occurs, there is little understanding of how best to intervene.

3. *Opportunities for academic and commercial research:*
   (i) New genome sequence knowledge. This would attract clinical trials of future DAAs to the UK, as the assessment of the emergence of resistance is a critical component of such studies. In addition, new determinants of treatment response will become apparent. Future clinical trials are more likely to be conducted in the UK in the event that full-length sequencing becomes readily available, with the potential to benefit UK wealth.
   (ii) Sequencing the human genome. Sequencing the human genome. There is a small amount of data on human genetic variation in relation to HCV infection. For example, polymorphisms at the IL28B locus can predict treatment response for genotype 1. However there remains unexplained non-response to interferon and DAA-based therapies. Further, the determinants of HCV-related hepatocelllar carcinoma and decompensated cirrhosis remain unclear. With the current (limited) state of knowledge, we consider that human genome sequencing in individuals with HCV infection represents an important area for research.

4. *Linkage with existing capacity and capabilities – key considerations for data workstream:*
   (i) An existing network of laboratories that perform viral gene sequencing. Currently, 17 NHS virology laboratories across England have the capability to sequence partial length HCV genomes because they already undertake partial sequencing of HIV as part of routine clinical care. Replacement of current testing platforms and protocols with newer technology would be associated with simplification of methodology. The cost per complete test would be anticipated to be comparable to existing costs (excluding the capital costs of instrument purchase). Only a few laboratories actually undertake HCV sequencing at present. A testing schedule will be required that defines when and how often testing is performed.
   (ii) A proposed mechanism for analysis of HCV gene sequences and reporting of drug resistance to clinicians. There are currently three linked initiatives that could be accessed to enhance the value of HCV whole genome sequencing: 1. HCV Research UK (MRF-funded). A network of 40 HCV clinics recruiting 10,000 HCV-infected patients across the UK with a common clinical database and biobank of samples. 2. STOP-HCV. A UK-wide research study funded by the MRC Stratified Medicine call to focus on viral and host determinants of HCV natural history and response to therapy, which incorporates most of the appropriate academic and clinical groups. 3. ICONIC. A Health Innovation Challenge Fund (WT/DH) programme to develop and optimise the methodologies for full-length viral sequencing. The analysis tools for producing a full-length consensus sequence and interpretation for clinicians are still to be developed. However, this will be addressed within the three existing initiatives described above, which will provide important proof of principle data. 4. HPA. A programme of partial HCV sequencing for drug resistance testing is underway.
   (iii) A proposed mechanism for curation of sequence data and linkage to clinical data. The existence of UK HCV Research UK (see above) provides an important opportunity to build a national clinical database to which viral gene sequences can be linked.

(iv) Validation of the phenotypic effect of novel gene mutations. The detection of a new mutation does not necessarily mean that this is associated with resistance, and requires validation. A number of research laboratories in the UK have the knowledge, expertise and capacity to undertake drug susceptibility testing using the HCV "replicon" systems, in order to test whether specific mutations do indeed lead to drug resistance.

5. *Key considerations for the ethics workstream*:
   Many of the same issues relate to HCV as for HIV gene sequencing. There is a far less well-developed patient advocacy environment, although organisations such as the British Liver Trust and the Hepatitis C Trust are actively raising awareness.
   (i) Identification of transmission chains. There is less awareness of the potential use of viral sequences to identify potential donors. Nevertheless, this could happen, and good governance of viral sequence data is essential.
   (ii) Availability of data for new interpretation. People having full-length viral sequencing undertaken will need reassurance that these data will be interrogated should a new viral genetic predictor of, for instance, disease progression, drug toxicities, or drug resistance be identified.

JUSTIFICATION FOR INCLUSION OF TUBERCULOSIS

1. *Burden of disease and existing need for sequencing:*
   (i) A total of 8,963 cases of tuberculosis were reported in the UK in 2011, a rate of 14.4 cases per 100,000 population. This is an increase on numbers in 2010. The majority of cases were notified from urban settings and affected young adults, notably those from countries with high TB burdens and those with social risk factors for TB. As in previous years, London accounted for the highest proportion of cases in the UK (39%), followed by the West Midlands (11%). Similarly to 2010, 74% of new cases were born outside of the UK and mainly originated from South Asia and sub-Saharan Africa. Rates in the UK born population remained stable at 4.1 per 100,000 population.
   (ii) The number of drug resistant cases continues to rise, with 431 cases (8.4%) resistant to any first line drug in 2011, up from 342 in 2010 – an increase of 26%. The number and proportion of isoniazid resistant and multi-drug resistant (MDR) cases also increased in 2011 (7.6% and 1.6%, respectively). Over the last decade, the proportion of MDR cases has increased from 0.9% in 2000 to 1.6% in 2011. The proportion of cases resistant to any first line drug was higher in those with a previous history of TB compared to those without, and in non-UK born cases compared to UK born. This pattern is similar for MDR cases. Twenty-four extensively drug resistant (XDR) cases have been reported in the UK since 1995, six of these in 2011.
   (iii) Two-thirds of notified cases of tuberculosis are confirmed by growth of the organism from patient samples. Other cases are treated on the basis of clinical history and examination, contact history, imaging and histological findings. Culture confirmation is lower in children, and in cases where the disease is outside the lung and requires a more invasive procedure to obtain samples for culture. The necessary laboratory support for the diagnosis and treatment of individuals with TB includes the assessment of infectivity and culture of samples in local NHS laboratories, followed by identification of *M. tuberculosis*, antimicrobial susceptibility testing and epidemiological typing in Health Protection Agency (HPA) reference laboratories. It typically takes two to three weeks to culture and identify *M. tuberculosis*, and between one to two months to achieve all of these goals, using a range of both traditional culture-based and molecular techniques. The development of molecular tests that simultaneously detect and identify *M. tuberculosis* and provide antimicrobial resistance data have improved the speed but not the sensitivity of TB diagnosis, and only target the most frequent genetic mutation for a limited number of first- and second-line antibiotics. Multidrug resistance and extremely drug resistant strains are uncommon in England but are of great public health importance.
   (iv) Whole genome sequencing is not used in clinical care, and is a research tool at the present time. Its introduction into routine care would revolutionise the way that TB is managed in relation to patient care and public health, and the potential gains are very high. Recently published and current translational research provides confidence of the potential for rapid progress in these areas.

2. *Expected impact of full-length bacterial sequencing on patient care and public health:*
   (i) Generating higher quality data. The implementation of next-generation sequencers in the care pathway of patients with TB would mean that full-length genome sequences would be generated. This would provide a full repertoire of genes to which the current anti-tuberculosis drugs are directed, and would allow rapid and accurate identification of *M. tuberculosis* and other mycobacterial species. Anticipated necessary genome coverage is 50x, which is easily achievable with current benchtop sequencers that can be used locally, and has been shown to deliver effectively a zero rate of false-positive single nucleotide polymorphism (mutation) calls on bacterial genomes.

(ii) Which patients would benefit. All newly diagnosed patients with culture-confirmed TB would be expected to benefit as soon as they enter the care pathway. This is tractable, with fewer than 10,000 new cases expected annually. New cases with resistance to one or more standard drugs would have more rapid identification of effective alternative drugs, which could reduce the period of infectivity of the patient, as well as reduce drug toxicity and improve outcomes. Mixed infections with multiple strains of TB will be detected more readily, with the potential to identify drug susceptibility for each strain. Known TB infected patients who required a change of therapy because of treatment failure and persistent culture-positivity would also be investigated using next-generation sequencing platforms, although the number of cases would be considerably lower than new cases (although of major importance, if drug resistance is the basis for treatment failure).

(iii) Generating data of public health importance. The incidence of tuberculosis is increasing within the UK. Public health control is currently supported by bacterial genotyping to determine whether two or more cases may be related by recent transmission. Since 2010, epidemiological typing has been performed by HPA reference laboratories on the first isolate from each new case of TB using a molecular method termed MIRU-VNTR. This is semi-automated and widely used in Europe and North America. It employs capillary sequencing or dHPLC (denaturing high-performance liquid chromatography). Clinical and public health teams use this information to prioritise the investigation of outbreaks and clusters, both locally and nationally, using a national database linked to the HPA clinical database. The limited resolution of the method may lead to false-positive links between cases, and the directionality and timing of transmission cannot be determined. Genome sequencing has a greater resolution than current typing methods, and will provide the opportunity to undertake phylogenetic analyses which would be predicted to bring a greater understanding of transmission patterns within and beyond England. Improved epidemiological typing could replace the existing methods and would be readily translated through existing public health systems. A number of phylogenetic methods are being developed by several research groups to exploit full-length *M. tuberculosis* sequences to describe the dynamics of spread, and to inform public health control.

3. *Opportunities for academic and commercial research:*

(i) New genome sequence knowledge. Full-length sequencing of *M. tuberculosis* would uncover new determinants of disease pathogenicity and drug susceptibility. This could be of utility to the pharmaceutical industry, with the potential for benefit to UK wealth. Current workflow patterns in which all isolates are referred to HPA reference laboratories provide an unusually robust collection of isolates, carefully linked to strain identity, drug susceptibility, epidemiological and clinical information, and form the basis for a strong national surveillance system.

(ii) Sequencing the human genome. Routine clinical TB care does not include host genome testing. We consider this to be a potentially important area of research.

4. *Linkage with existing capacity and capabilities – key considerations for data workstream:*

(i) Mycobacterium Reference Laboratories. There are three TB reference laboratories in England (Newcastle, Birmingham and London), serving relevant geographical areas in the North, Midlands and South. These provide a major role in TB diagnosis, undertake tests to confirm the identity of suspected *M. tuberculosis,* perform drug susceptibility testing and undertake epidemiological typing by 24 locus MIRU-VNTR. All of the information reported to NHS service users contributes invaluable data to national surveillance. Analysis of clusters is carried out locally and nationally, and links typing data with clinical and risk factor information. Whole genome sequencing is not part of routine testing, and is currently a research tool only. A project in Oxford and Birmingham to implement TB (microbial) whole genome sequencing for individual patient care, 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).

(ii) Laboratories with the capability to perform *M. tuberculosis* whole genome sequencing. Although the number of laboratories that perform *M. tuberculosis* sequencing is currently very small, any laboratory that has a sequencing platform could undertake this, working with clinical laboratories that culture *M. tuberculosis*. Areas with high incidence of TB are mainly urban settings with access to academic centres or large clinical diagnostic laboratories where both isolation and whole genome sequencing could be carried out, reducing avoidable delays. Sequencing of *M. tuberculosis* could be undertaken within the network of laboratories that currently perform, or will later adopt viral sequencing for HIV/HCV. This would be efficient, and would reduce the turnaround time to *M. tuberculosis* sequence data becoming available. We recommend this devolved model of genome sequencing rather than sole reliance of a small number of reference laboratories, but caution that laboratories carrying out this work accept responsibility for reporting to clinicians, public health networks and local and national surveillance systems, and submission of sequences to relevant repositories. Referral systems must remain in place

for smaller services in low incidence areas, and be kept under review during the reconfiguration that will result from Pathology Modernisation.

(iii) Genome sequence repository. Several centres have generated *M. tuberculosis* whole genome databases, but these are not shared with the exception of data generated by the Sanger, whose data are placed into a publically accessible short read archive. A single database will be required for national surveillance and control of TB. Deposition of *M. tuberculosis* gene sequence generated for clinical purposes in England by any diagnostic laboratory used by the NHS should be mandatory, and the necessary control should be established to separate clinical and patient identifiable information (which is required for public health management), from sequence data (which could be shared more widely).

(iv) Analysis of *M. tuberculosis* gene sequence to confirm bacterial identity and predict drug susceptibility. Whole genome sequence could provide an accurate method of bacterial identification, and would obviate the need for alternative identification methods. Some work to validate the process of identification for non-tuberculous *Mycobacterium* species would also be useful for particular groups of vulnerable patients, such as those with cystic fibrosis.

(v) 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 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. We recommend that phenotypic testing, including susceptibility testing, remain within the reference laboratory system to maintain expertise. However, there should be sufficient flexibility in the system for new research findings described by any researcher to be examined, and added to the catalogue of known resistance mechanisms if verified. The detection of mutations associated with known resistance will require automation through the development of new bioinformatics tools. We recommend that these be available to all laboratories that deposit whole genome sequence data into the national database, since this will allow any laboratory to interpret their sequence data in relation to known resistance mutations.

(vi) Analysis of *M. tuberculosis* gene sequence to investigate source of origin and outbreaks. The first isolate from all new cases of TB diagnosed in England should undergo whole genome sequencing. This will replace the current genotyping method (VIRU-VNTR), which is less discriminatory. Using the database discussed above, each new isolate would be 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. *M. tuberculosis* isolates obtained during the course of an outbreak investigation would be compared with each other and with the database to assess whether recent transmission has occurred. 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.

(vii) Linkage between genome sequence and clinical data. 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 TB. We recommend that this be developed, and that this function be aligned with the genome repository function.

5. *Key considerations for the ethics workstream*:

(i) Identification of transmission events. When a new case of TB is diagnosed in the UK, it is standard practice to investigate whether an infected source can be identified, and whether other contacts from the household, workplace or other shared areas have also been affected. Although the availability of whole genome sequence data will increase the genetic discrimination between strains of *M. tuberculosis* and thereby increase the accuracy with which transmission events can be defined, the standard process of outbreak investigation is unlikely to change. The accurate interpretation of sequence data will depend on careful consideration of the epidemiological information, but decisions on whether to extend investigations can be prioritised by improved understanding of the speed and extent of transmission. An important point for consideration is the management and sharing of patient identifiable data, clinical details and contact details, which will need to be linked to genomic data for maximum public health

benefit. Currently, this information is held by the HPA who generate all of the typing information and perform cluster analyses, working with clinical colleagues who manage the patients. If genomic data becomes generated by a much wider network of laboratories and the genomic database is shared more widely, the information governance for data flows will need to be carefully considered.

Faculty of Medical Sciences, University College London, Gower Street , London, WC1E 6BT.
Tel: +44 (0) 20 7679 2000

Appendix 4. Membership. CMO Genomics science priorities advisory group

Professor David Lomas (Chair) – Dean of the Faculty of Medical Sciences and Chair of Medicine, UCL

Dr Helen Firth - Consultant Clinical Geneticist, Cambridge University Hospitals Trust

Professor Herbie Newell – Professor of Cancer Therapeutics - Northern Institute for Cancer Research

Dr Julian Parkhill – Head of Pathogen Genomics, The Sanger Institute

Professor Sharon Peacock (Lead, working group infectious diseases) – Professor of Clinical Microbiology, Department of Medicine, University of Cambridge

Professor Deenan Pillay – Head of the Research Department of Infection and Honorary Consultant Virologist, UCL

Dr Grace Smith –Consultant Medical Microbiologist, HPA Regional TB Reference Laboratory, Heart of England Hospital, Birmingham

Professor Brian Spratt – Wellcome Trust Principal Research Fellow, Imperial College London

Dr John Bradley – Director, NIHR Cambridge Biomedical Research Centre

Professor Patrick Chinnery (Lead, working group rare diseases) – Wellcome Trust Senior Fellow in Clinical Science, Professor of Neurogenetics and Honorary Consultant, University of Newcastle

Professor Sian Ellard – Professor of Human Molecular Genetics, University of Exeter Medical School and Consultant Molecular Geneticist, Royal Devon & Exeter NHS Foundation Trust

Professor Mark McCarthy – Robert Turner Professor of Diabetic Medicine, University of Oxford

Dr Ultan McDermott – Career Development Fellow Group Leader in the Cancer Genome Project, Wellcome Trust Sanger Institute

Professor David Cameron (Lead, working group cancer) – Professor of Oncology at Edinburgh University and Director of Cancer Services in NHS Lothian

Professor Andrew Hanby – Chair In Breast Cancer Pathology, University of Leeds

Professor Charles Swanton – Chair in Personalised Cancer Medicine, UCL

Professor David Livermore – Professor n Medical Microbiology, University of East Anglia

Dr Ian Walker – Head of Stratified Medicine and Combinations Alliance, Cancer Research UK

Mark Bale – Interim Director, Health Science and Bioethics Division, Department of Health

Secretariat:

Hilary Tovey – Senior Policy Manager, Cancer Research UK

Dr Cathleen Schulte – Senior Policy Lead, Health Science and Bioethics Division, Department of Health

Colin Pavelin – Head of Genomics and Rare Diseases, Health Science and Bioethics Division, Department of Health

Daniel Bridge – Policy Adviser, Cancer Research UK