wellcome trust sanger institute

Wellcome Genome Campus
Hinxton Cambridge CB10 1SA
T 01223 834244
F 01223 494919
www.sanger.ac.uk

22nd June 2016

RE: Notice Number: NOT-RM-16-025
Wellcome Trust Sanger Institute Response to NIH Request for Information: Characterizing and Understanding the Organization of Individual Cells within Human Tissues.

Introduction

The Wellcome Trust Sanger Institute (WTSI, www.sanger.ac.uk) is a non-profit, British genomics and biodata research institute primarily funded by the Wellcome Trust. There are two molecular atlas projects based at the WTSI that are relevant to this NIH RFI call, as well as associated relevant technology development activities. We will describe the atlas projects in section 1. below, and outline the areas of technology development in section 2. below.

1. Existing molecular atlas projects at the WTSI

At the WTSI there are two major molecular atlas projects that are relevant to this Request for Information: a "Human Cell Atlas" and "Human cell lineages from somatic mutations".

(i) Human Cell Atlas

The first project is a census of all human cell types profiled by single cell RNA-sequencing, dubbed the "Human Cell Atlas". The conceptual framework of this project is that characterization of cell types remains surprisingly limited to this day, and has been acquired to a large extent through a combination of microscopy and focus on a limited number of marker genes/proteins. Genomics offers the promise of a systematic approach. However, so far, it has been applied largely in bulk to ensembles of hundreds or thousands of cells — thus masking critical differences between individual cells. Exciting recent advances in high-throughput single cell genomics have suddenly put a systematic, high-resolution Human Cell Atlas within reach.

Clearly, this is a tremendously ambitious project which can only be achieved through a globally coordinated and collaborative effort. Dr Sarah Teichmann (Head of Cellular Genetics, WTSI) is actively setting up an international network, together with Dr Aviv Regev (Chair of Faculty, Broad Institute) and the support of the Wellcome Trust staff members, including Michael Dunn (Head of Genetics and Molecular Biosciences, Wellcome Trust, London). The initial steps of this coordination have included setting up a steering committee with representatives from USA/UK/Sweden/Japan/Israel and elsewhere, and organising a kick-off meeting at the Wellcome Trust headquarters in London on October 13-14th 2016.

It is likely that this global effort will be organised along the lines of previous international consortia in genomics, such as the International Cancer Genome Consortium (ICGC). Key points are adherence to a common set of protocols (both experimental and computational), focus on collaborative not competitive research, agreed data sharing and ethical standards, and a joint data portal. In addition, it will be necessary to agree on a reference list of human tissues and organs, drawing on the knowledge in the community from sources such as pathologists and databases such as cellPedia at http://shogoin.stemcellinformatics.org/. For each tissue, the target will be to apply an agreed set of genomics protocols and sequence a minimum number of high quality cells (e.g. several thousand).

There are clearly challenges to obtaining high quality cDNA from human cells that are representative

wellcome trust sanger institute

Wellcome Genome Campus
Hinxton Cambridge CB10 1SA
T 01223 834244
F 01223 494919
www.sanger.ac.uk

of the in vivo state, with cells responding to disaggregation by stress response and apoptosis. At the same time, there are papers that have successfully disaggregated adherent cells and drawn insightful new biological conclusions, e.g. interactions between cells in melanoma (Tirosh *et al.*, 2016) and new intestinal cell types (Grun *et al.*, 2016). Furthermore, there are technical solutions that have allowed gentle disaggregation and cell capture of mouse neurons which have led to discoveries of new neuronal cell types (Zeisel et al., 2015) and of mouse retinal cells (Macosko *et al.*, 2015), both adherent cell types. Sten Linnarsson leads the single cell genomics facility at the Karolinska Institute (Stockholm, Sweden), and as a consequence this group has accumulated extensive expertise in disaggregation, lysis and capture of cells from many different tissues. Furthermore, there are now working protocols for nuclear RNA-sequencing (Habib *et al.*, biorxiv), offering an alternative to cytoplasmic RNA capture.

On a practical level, the human material used in pilot projects as part of the "Human Cell Atlas" is currently obtained either from biopsies as a by-product of clinical diagnostic tests (e.g. lung biopsies for analysis of potential asthmatic individuals) or from organ donors where the organ was deemed unsuitable for transplantation e.g. due to a ruptured blood vessel (rather than evidence of disease). These are the sources which provide as close as possible to "healthy" donor tissue. In future, it may be possible to recruit tissues in a systematic way from individual donors along the same lines as the GTex consortium.

Looking even further forwards, additional dimensions beyond a healthy reference adult human cell atlas will be mapping of disease states of tissues and organs, describing developmental states and studying iPSC-derived cell types.

In summary, this project will transform biology by leading to the discovery of a rich diversity of hitherto unappreciated cell states and cell types.

ii) Human cell lineages from somatic mutations

While the "Human Cell Atlas" project described above aims to chart cell types and cell states, there will not be explicit lineage information inherent in this project. Extracting information about lineage, and simultaneously collecting data on the mutational burden of cells, is the aim of the "Somatic mutation" project at the WTSI based on single cell DNA-sequencing technology.

At present, we have limited insight into the mutational burden and signatures of human cells during normal development and physiological adult cell turnover, nor how this might vary in developmental disorders, with ageing or across a range of diseases. Studies of blood samples obtained from haematologically normal individuals have shown that large clones carrying somatically acquired driver mutations are surprisingly prevalent in the middle-aged to elderly population (Busque *et al.*, 2012; Jacobs *et al.*, 2012; Laurie *et al.*, 2012). We have reported that normal, sun-exposed skin cells carry a heavy burden of somatic mutations, with up to 30% of cells carrying one or more *bona fide* driver mutations that confer a clonal advantage on the cell (Martincorena *et al.*, 2015). Somatically acquired large-scale copy number variants have been identified in a variety of normal tissues, including in significant proportions of human neurons, lymphoblasts and fibroblasts (Abyzov *et al.*, 2012; Cai *et al.*, 2014; McConnell *et al.*, 2013; O'Huallachain *et al.*, 2012). The evidence from these studies is that histologically normal tissues can carry significant mutational burden, including oncogenic variants, and yet fulfil their required functions.

Furthermore, the data illustrated in the Figure below (Alexandrov *et al.*, 2015), shows that tissues right across human physiology all accumulate somatic mutations. A key point is that these increase throughout human life in a clock-like manner with particular mutational patterns (or "signatures").

wellcome trust sanger institute

Wellcome Genome Campus
Hinxton Cambridge CB10 1SA
T 01223 834244
F 01223 494919
www.sanger.ac.uk

This mutational burden, with distinct sequence biases in different tissues, may thus be a fundamental part of the phenotype of each cell type. Integrating this information with the transcriptomic signature from the Human Cell Atlas may yield interesting and unexpected insights into cell physiology.

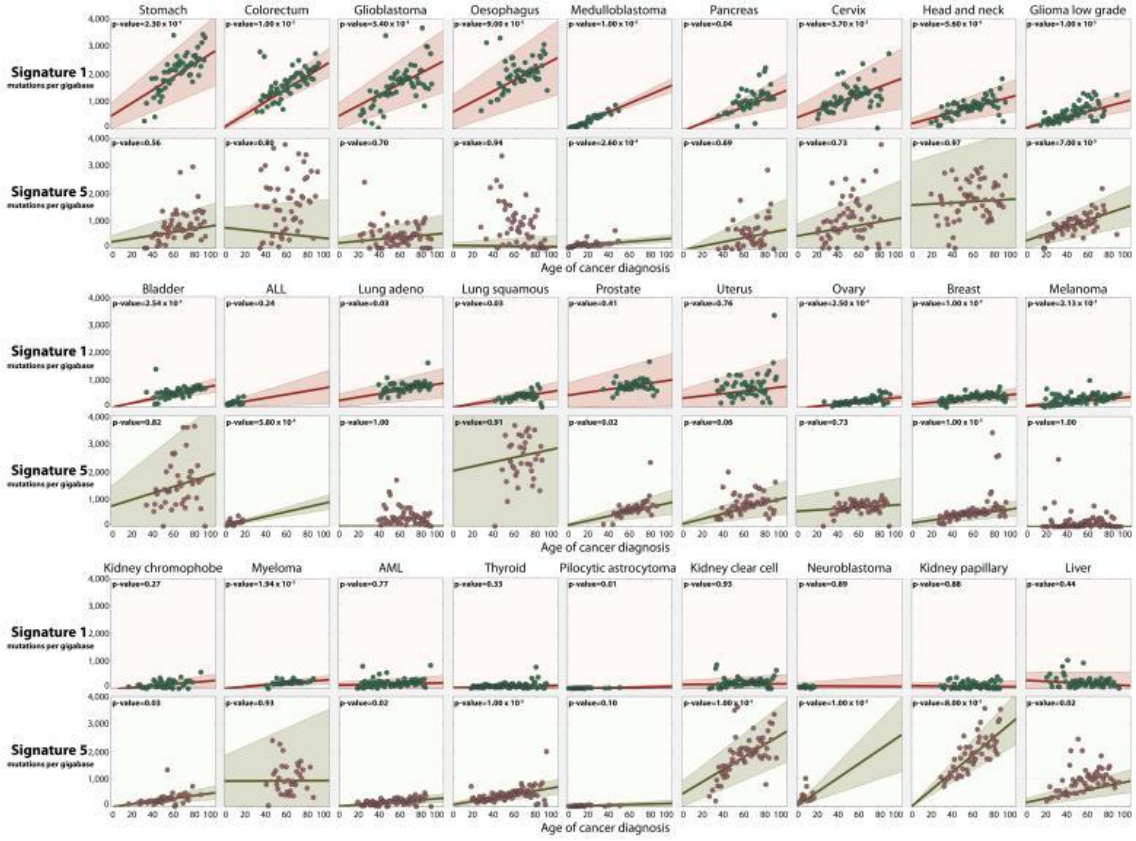


Figure. Correlations between ages of cancer diagnosis and mutations attributed to mutational signatures (1 and 5) f (Alexandrov *et al.*, 2015)

Y-axes show numbers of somatic substitutions per gigabase attributed to either mutational signature 1 or signature 5, while X-axes show ages of cancer diagnosis. Each panel corresponds to a cancer type and panels are sorted in a decreasing order of the estimated slopes for signature 1. Each dot represents the median number of somatic mutations for all cancers of a given age. Red and green lines show best estimates for the slopes, *i.e.*, mutation rates, of signature 1 and 5 respectively. 95% confidence intervals for the slopes are shown in lighter green and lighter red shading for signatures 1 and 5 respectively. Note that for several cancer types, slopes extend far beyond the available data points; this representation is not intended to be a prediction but rather it is done for consistent presentation across all panels in the figure.

The information inherent in the somatic mutations can also be harnessed to reconstruct recent clonal relationships between cells as well as more distant developmental relationships. Using shallow, single cell whole genome sequencing coupled with high-throughput genotyping methods, we can reconstruct cellular lineages to the depth of the first few thousands of cells of the embryo from mutation catalogues in adult cells. We have published a proof-of-principle study in two adult mice (Behjati *et al.*, 2014) and have equivalent unpublished data in haematopoietic cells from an adult human. For a human cell lineage project, the source of material would ideally be broad biopsy samples from single human individuals, through so-called "warm autopsies" of consented donors.

These studies will enable us to investigate:

wellcome trust sanger institute

Wellcome Genome Campus
Hinxton Cambridge CB10 1SA
T 01223 834244
F 01223 494919
www.sanger.ac.uk

- The contributions of cells at different numbers of cell generations from the fertilised egg to each differentiated cell type and hence the temporal transitions during early embryonic divisions from pluripotency to multipotency to oligopotency to unipotency;
- The relationships between lineages of different cell types;
- The cellular ancestry of physically proximate and distant cells revealing potential cell migration routes, the development of laterality and cellular ancestries from the perspective of the axial body plan;
- Functional plasticity across clades of phylogenetically related cells.

2. Technologies

The two projects described above are underpinned by extensive genomic technology development at the level of single cell genomics and computational approaches, which we will outline in the next two sections below.

(i) Single cell genomics technologies

The WTSI has invested heavily in both developing as well as scaling up single cell genomics protocols. We have used both commercial instruments (*e.g.* Fluidigm C1, 10x Genomics Chromium) and in-house research solutions. Examples of an in-house systems are FACS index sorting of single cells coupled with plate-based liquid handling robotics, as well as droplet microfluidics for cell capture and molecular biology. The protocols that are "production-ready" they are ported to the Single Cell Genomics Platform, where staff members offer them as a high-throughput pipeline service to the institute, which is key to achieving the large projects mentioned in section 1 above:
http://www.sanger.ac.uk/science/groups/single-cell-genomics-core-facility

Many of the principle investigators involved in developing single cell genomics and single cell bioinformatics technologies are part of the Sanger-EBI Single Cell Genomics Centre, which was founded in 2012/13 by five PIs (Marioni, Ponting, Reik, Teichmann, Voet) from the WTSI and the EMBL-European Bioinformatics Institute, which are co-located institutes on the Wellcome Genome Campus outside Cambridge/UK:
http://www.sanger.ac.uk/science/collaboration/sanger-institute-ebi-single-cell-genomics-centre

Amongst the first activities of the Sanger-EBI Single Cell Genomics Centre were the development of a scDNA-seq protocol by Thierry Voet and colleagues (Voet *et al.*, 2013) as well as the assessment of scRNA-seq protocols using external spike-in RNA and innovative statistical methods (Brennecke *et al.*, 2013). Both development of high accuracy single cell DNA-sequencing protocols and comparisons of scRNA-seq protocols based on spike-ins to assess sensitivity and specificity continue to be ongoing research activities in the centre.

scDNA-seq and scRNA-seq are tremendously powerful standalone techniques, and combining them together in individual cells holds huge promise. Multi-omics single cell methods will enable the interrogation of (epi)genomic differences across the cells and reveal how this correlates with transcriptomic variation. The epigenome is fascinating from the point of view of regulation, while the DNA sequence provides data on mutational burdens and clonality as outlined above.

The Voet group, together with Reik and Ponting has been at the cutting edge of multi-omics methods development. For example the G&T-seq protocol applies robotically controlled physical separation

wellcome trust sanger institute

Wellcome Genome Campus
Hinxton Cambridge CB10 1SA
T 01223 834244
F 01223 494919
www.sanger.ac.uk

of the polyadenylated RNA from the DNA of a lysed single cell using oligo-dT coated magnetic beads with subsequent amplification and preparation of the RNA- and DNA-seq libraries in parallel (Macaulay *et al.*, 2015). G&T-seq was further developed to M&T-Seq (Angermueller *et al.*, 2016), enabling DNA-methylome and transcriptome profiling of the same single cell by establishing DNA bisulfite- and RNA-seq of the same cell in parallel. Further development of multi-omics single cell methods continues to be an active area of research by Voet and others at the WTSI.

All the single cell genomics protocols mentioned above rely on cell disaggregation and cell or nuclear isolation, without taking into account the context of the cell within the tissue. Maintaining or adding information about cell neighbourhoods and cell-cell contacts and interactions through spatial information profoundly enriches the genomic data. The Marioni group has driven the computational integration of systematic in situ hybridisation data with single cell RNA-seq data (Pettit *et al.*,2014; Achim *et al.*, 2015). Avenues for systematic spatial positioning of cells in tissues are integrative techniques using in situ hybridisation or immunohistochemistry, or tissue mass cytometry (Giesen *et al.*, 2014) combined with single cell genomics, or direct in situ sequencing (e.g. FIS-SEQ, Lee *et al.*, 2014), or imaging methods (MERFISH, Chen *et al.*, 2015). A further way of achieving genomics with spatial information is to capture individual cells and register their spatial context at the same time, as is possible by laser capture microdissection for instance.

(ii) Computational approaches to single cell genomics data analysis

All of the technologies and projects mentioned above are drive by computational methods development for data processing and interpretation. There is intense interest in this area on the Wellcome Genome Campus from the Brazma, Hemberg, Marioni, Stegle and Teichmann groups as well as others, and software and databases are available at:
www.singlecellbioinformatics.org

Computational methods include normalizing and processing single-cell genome and transcriptome datasets (Brennecke *et al.*, 2013; Buettner *et al.*, 2015; Ilicic *et al.*, 2015; Lun *et al.*, 2016) and developing automatic quality control criteria to exclude broken cells or otherwise unsuitable samples within datasets (Ilicic *et al.*, 2016). The next level of analysis includes modelling technical and biological factors (e.g. cell cycle, "Cyclone" software) that drive single cell heterogeneity within samples. The Marioni and Stegle groups have developed statistical and machine learning methods to regress and quantify these factors underlying transcriptomic cell-to-cell variation (Buettner *et al.*, 2015; Vallejos *et al.*, 2015; Vallejos *et al.*, 2016).. The Hemberg group is developing next generation clustering methods to assign cell types and cell states in an unbiased manner using high-dimensional statistical models, and to extract marker genes that correlate uniquely with clusters.

Overall, it is critical that computational methods scale with the volume of data, and can cope with integration of data from diverse experiments and ideally also diverse genomics protocols by modelling batch effects. Finally, suitable databases will need to be developed that serve and visualize single cell genomics data in appropriate ways. The Brazma group, together with Marioni and Teichmann, are funded to develop a "Single cell Gene Expression Atlas" as a next generation of their "Gene Expression Atlas" linked to ArrayExpress:
https://www.ebi.ac.uk/gxa/home
An example of a single cell gene expression database is "Espresso", developed by the Teichmann group for their mES cell scRNA-seq datasets:
http://www.ebi.ac.uk/teichmann-srv/espresso/

wellcome trust sanger institute

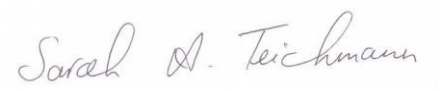
Wellcome Genome Campus
Hinxton Cambridge CB10 1SA
T 01223 834244
F 01223 494919
www.sanger.ac.uk

In summary, the time is now right for large-scale, quantitative approaches to studying human cells in their physiological context based on single cell genomics and related methods. We are enthusiastically embracing these opportunities on the Wellcome Genome Campus and hope to synergize with NIH initiatives in this area.

Signed:

Professor Sir Mike Stratton FMedSci FRS
Director, Wellcome Trust Sanger Institute
Chief Executive, Wellcome Genome Campus

Dr Sarah Teichmann, FMedSci
Head of the Cellular Genetics Programme
Wellcome Trust Sanger Institute



Dr Peter Campbell
Head of the Cellular Genetics Programme
Wellcome Trust Sanger Institute

Dr Thierry Voet
Group Leader, WTSI
Assistant Professor, KU Leuven

Dr John Marioni
Associate Faculty, WTSI
Group Leader, EMBL-EBI
Senior group leader, CRUK Cambridge Institute

Dr Martin Hemberg
CDF Group Leader
Wellcome Trust Sanger Institute

Dr Oliver Stegle
Group Leader
EMBL-European Bioinformatics Institute

wellcome trust sanger institute

Wellcome Genome Campus
Hinxton Cambridge CB10 1SA
T 01223 834244
F 01223 494919
www.sanger.ac.uk

References

Abyzov, A. et al. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature 492, 438–442.

Achim, K. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin.Nat Biotechnol 33(5):503-9.

Alexandrov, LB. et al. (2015). Clock-like mutational processes in human somatic cells. Nat Genet. 2015 Dec;47(12):1402-7.

Angermueller, C. et al. (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nature methods 13, 229-232

Behjati, S. et al. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature 513, 422–425.

Brennecke, P. et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 10(11):1093-5.

Buettner, F. et al. (2015) . Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Bioetchnol. 33(2):155-60

Busque, L. et al. (2012). Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. Nat Genet 44, 1179–1181.

Cai, X. et al. (2014). Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep 8, 1280–1289.

Chen KH. et al. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348(6233):aaa6090.

Giesen, C. et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat Methods. 2014 Apr;11(4):417-22.

Grün, D. et al. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015 Sep 10;525(7568):251-5.

Habib, N. Div-Seq: A single nucleus RNA-Seq method reveals dynamics of rare adult newborn neurons in the CNS.

Ilicic, T. et al. (2015). Classification of low quality cells from single-cell RNA-seq data. Genome Biol 17;17:29.

Jacobs, K.B. et al. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet 44, 651–658.

Kim, JK. et al. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat Comm 22;6:8687.

Laurie, C.C. et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet 44, 642–650.

Lee JH. et al. (2014). Highly multiplexed subcellular RNA sequencing in situ. Science ;343(6177):1360-3.

Lun, A. et al. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol DOI: 10.1186/s13059-016-0947-7

Macaulay, I. C. et al. (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nature methods 12, 519-522.

wellcome trust sanger institute

Wellcome Genome Campus
Hinxton Cambridge CB10 1SA
T 01223 834244
F 01223 494919
www.sanger.ac.uk

Macosko, EZ. et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161(5):1202-14.

Martincorena, I. et al. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. Science (80-) 348, 880–886.

McConnell, M.J. et al. (2013). Mosaic copy number variation in human neurons. Science 342, 632–637.

O'Huallachain, M. et al. (2012). Extensive genetic variation in somatic human tissues. Proc Natl Acad Sci U S A 109, 18018–18023.

Pettit, JB. et al. (2014). Identifying cell types from spatially referenced single-cell expression datasets. PLoS Comp Biol 10(9):e1003824.

Tirosh, I. et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352(6282):189-96.

Vallejos, CA. et al. (2015). BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLoS Comp Biol 11(6):e1004333.

Vallejos, CA. et al. (2016).Beyond comparisons of means: understanding changes in gene expression at the single-cell level. Genome Biol 15;17(1):70.

Voet, T. et al. (2013). Single-cell paired-end genome sequencing reveals structural variation per cell cycle. Nucleic Acids Res 41(12):6119-38

Zeisel, A. et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.Science. 2015 Mar 6;347(6226):1138-42.