A107

The molecular basis of the cellular taxonomy of the human body

Alessandra Breschi[1,2], Carrie Davis[3], Sarah Djebali[4], Jesse Gillis[3], Dmitri D. Pervouchine[5], Anna Vlasova[1], Alex Dobin[3], Chris Zaleski[3], Jorg Drenkow[3], Cassidy Danyko[3], Alexandra Scavelli[3], Manuel Munoz[1], Diego Garrido[1,2], Ferran Reverter[1], Thomas R. Gingeras[3], Roderic Guigo[1,2]
[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain; [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain; [3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11742, USA; [4]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, Castanet Tolosan, France
Human Genomics 2018, **12(Suppl 1):**A107

The collection of RNA sequencing experiments produced by the ENCODE project from many cell lines and bulk tissues constitutes a comprehensive catalogue of the expression programs utilized by human cells. However, the relationship between the transcriptomes of tissues and the transcriptomes of the constituent primary cells, and how these impact tissue phenotypes has not been well established. Here we have produced RNA sequencing data for a number of primary cells from multiple human body locations. The analysis of this data, together with additional epigenetic data also produced by the ENCODE project for a total of 146 primary cells, indicate that most cells in the human body belong to five major cell types: epithelial, endothelial, mesenchymal, neural and blood cells. These redefine, based on gene expression, the basic histological types in which tissues are usually classified. We identified genes specific to these cell types, including a core set of transcription factors (TFs). Cell type specific genes, particularly when lying in open chromatin domains, are enriched for motifs for these cell type specific TFs, suggesting that they are potential candidates to drive cell type specificity. We estimated the relative proportion in tissues of the different cell types based on the transcriptional profiles obtained from bulk tissue sections. This inferred cellular composition is a characteristic signature of tissues and reflects.

A108

Know thy cells: the classification of tumor models by tissue, disease, sex, and species

Heather Selby[1], Mark Kon[1,2], John Quackenbush[3,4,$], and Benjamin Haibe-Kains[5,6,7,8,$]
[1]Boston University, Bioinformatics Program, Boston, MA, USA; [2]Boston University, Mathematics and Statistics, Boston, MA, USA; [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA USA; [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA USA; [5]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada; [6]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada; [7]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada; [8]Ontario Institute of Cancer Research, Toronto, Ontario, Canada
Correspondence: Heather Selby
Human Genomics 2018, **12(Suppl 1):**A108

Cell lines and xenografts are models of patient tumors. Since the first patient-derived tumor cell line, HeLa, was created, new cell lines have been repeatedly established from patients. Large cell line collections now broadly capture the genomic diversity of patient tumors, and provide valuable insight into drug response. Patient-derived xenografts (PDXs) have begun to replace cell lines as more predictive experimental tumor models for drug development. PDXs are directly established from patient tumors in immune-deficient mice, and better resemble the actual patient tumor than cell lines. Recently, the U.S. National Cancer Institute (NCI) retired their NCI-60 panel of patient-derived tumor cell lines to refocus their anti-cancer drug screening on PDXs.

Misidentified cell lines and xenografts are invalid models of patient tumors. The growth of cells in culture or xenografts makes tumor models susceptible to cross-contamination and/or mislabeling in spite of best laboratory practice. Although cell line misidentification is now widely-recognized, and authentication by short tandem repeat (STR) recommended, misidentified cell lines continue to be pervasive problems in cancer research. The International Cell Line Authentication Committee (ICLAC) has documented as many as 451 misidentified cell lines. Misidentified tumor models are often not from their original donor, and may even come from another tissue, disease, sex, or species.

To address the problem of misidentified tumor models, we developed a new classifier, Top Scoring-Binary Gene Pair (TS-BGP). The TS-BGP classifier, trained on tumor gene expression profiles, was used to predict the tissue, disease, sex, and species of both cell lines and PDXs. Binary gene pair signatures were extracted from the 31 tissues in the Genotype-Tissue Expression (GTEx) project, 32 diseases in The Cancer Genome Atlas (TCGA), 2 sexes in the GTEx project, and 2 species in the Encyclopedia of DNA Elements (ENCODE). The predictive accuracies of the leave-one-out-cross validations were 97.51% (GTEx; n=9115) for tissue, 90.89% (TCGA; n=10088) for disease, and 99.10% for sex (GTEx; n=9115), and 100% for species (ENCODE; n=622).

The TS-BGP classifier identifies tumor models that accurately resemble patient tumors. With tumors models that resemble patient tumors, data generated in pre-clinical research can successfully identify translatable drugs and therapies that demonstrate clinical efficacy. The TS-BGP classification of tumor models by tissue, disease, sex, and species helps ensure the credibility, reproducibility, and translation of pre-clinical research.

A109

RIKEN Ageing Resource Data Project: Single cell transcriptome and epigenetic changes in mice

Tommy Terooatea[1,4], Tsukasa Kouno[1], Matteo Guerrini[2], Yasutaka Motomura[2], Naoko Sato[2], Takeshi Matsui[2], Sidonia Fagarasan[2], Hironobu Fujiwara[3], Kazuyo Moro[2], Hiroshi Ohno[2], Ichiro Taniuchi[2], Kosuke Hashimoto[1], Piero Carninci[1], Aki Minoda[1,4]
[1]RIKEN CLST-DGT, Yokohama, Japan; [2]RIKEN IMS, Yokohama, Japan; [3]RIKEN CDB, Kobe, Japan; [4]Department of Internal Medicine, School of Medicine, Keio University, Tokyo, Japan
Correspondence: Tommy Terooatea; Aki Minoda
Human Genomics 2018, **12(Suppl 1):**A109

Bulk analyses of high-throughput genomic and proteomic technologies have produced invaluable publicly available data on ageing. However, these data are limited in terms of accurately defining cellular states, development and disease state. Single-cell omics on the other hand help overcome these obstacles and promises greater understanding of cellular processes. Consequently, we are currently generating a **resource dataset for ageing studies** with a focus on single cell RNA-seq. We present here our initial results produced from various tissues taken at different ages of the mouse. One aspect of our analysis will be to determine whether we observe dynamic evolution of cells in heterogeneity throughout ageing, which may reveal key regulatory pathways that are gradually deregulated during ageing.

We will also present our plan for obtaining single cell ATAC-seq and multi-omic datasets (DNA methylation, CAGE, translatome (ribosome profiling), proteomics and metabolomics) for the selected cell types of interest and integration analysis. Additionally, a unique aspect of our dataset is the utilization of both germ-free and SPF mouse models, which will enable us to introduce and access the effects of