Yu, Next Generat Sequenc & Applic 2015, 2:2
DOI: 10.4172/2469-9853.1000e104

Journal of
Next Generation Sequencing & Applications

Editorial Open Access

# The Human Transcriptomes Project: Is it hard?

**Jun Yu***

*CAS Key Laboratory of Genome Science and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.*

## Editorial

While cheering 10-thousand, 100-thousand, and one-million human medical sequencing projects happening here or there for precision medicine, we should still be worrying about the next, the long-expected, large-scale biology initiative – the Human Transcriptomes Project. Although we have enjoyed large-scale data for human transcriptomes since early 1990s, generated from EST (expressed sequence tag), microarray, and NGS (next generation sequencing), the first *draft human transcriptomes* have yet to be celebrated, let alone any of the complete. Why is this?

The human transcriptomes – the ribogenome of the human body – are hard to define, which are composed of RNA transcripts of all given cell types. The transcripts can be as large as hundreds of kilonucleotides, as small as tens of nucleotides, and as more as one half to one million per cell, and thus categorizing them all for the human body or even for a given cell type is an enormous task. That is not yet all. First, the human body is said to have $10^{13}$ physiologically dynamic cells, partitioned into a few tens of tissue types and a few hundreds of cell types, which develop, differentiate, specialize, age, and die voluntarily or passively. Second, cell is a functional unit of life, especially multicellular organisms, which not only has its life span and propagation cycle but also resides in a niche that nurtures itself and neighbors – not necessarily in contact – to whom it communicates with. Third, the human body has a life span that is highly variable longitudinally in size and physiology, such as fertilized egg, embryo, infancy, puberty, menopause, etc. Fourth, transcriptomes also vary among individuals since genomes and epigenomes are both variable albeit often in dramatically different degrees; genome variation is easy to define, such as based on high-coverage genome sequences, but epigenome variation is difficult to be identified since it is neither passed on to offspring faithfully nor stable within an individual of a stable genetic background.

How to define a human transcriptome? First, the function of a given transcript follows one of the two tracks: operational or informational; the former includes all RNAs and the latter refers to only the protein-coding part of massager RNAs or rather the translation (to protein) and the change of such part. Second, the expression level of each transcript has to be stratified into variable and invariable among the cell and tissue types; the former is expected to be limited in number and the latter covers most of the transcripts. The dynamic range of the transcripts covers several orders of magnitudes: from a single copy to 100,000s copies. Third, the expression patterns are expected to be more complex than what we have realized now. We perhaps have to borrow some concepts from microbiology where both a relatively stable *core transcriptome* and a highly variable *transcript cloud* have to be defined under certain condition for each cell type. The variability may come from individual person's or cell type's genetic background or epigenomic status, physiological condition, and experimental limitation. Across diverse tissue or cell types, universally-expressed and tissue-/cell-specific genes have to be defined in a hierarchical structure regardless their expression level variability, and so do physiology- and function-associated genes. Fourth, disease starts with functional failure of cells; cells under various pathological conditions are to be classified and their transcriptomes are disease-associated. The comparison cell-specific transcriptomes between the normal and the ill are of essence for both diagnostics and treatment.

The technology for a large-scale collaborative transcriptomic study is not yet ready but expected to be ready within the early phase of such a project. The weakness of our current technology for transcriptomics is also multi-fold. Although the current workhorses, such as the platforms offered by Illumina and Thermo-Fisher Scientific, very much satisfy the throughput need, the length-related quality is still a serious issue, when mapping transcripts involved in alternative starts, alternative exon-splicing and alternative polyadenylation, not mentioning large families of paralogous genes. More challenging breakthroughs are waiting for Great Leap Forward, involving at least three basic parameters: longer read-length beyond one kilobases, down-to-earth identification of RNA modification, and direct sequencing of RNA molecules without cDNA intermediates. We are not seeking for one-stop solutions for all but some and better. Another major concern is per cell resolution. While hooraying for single-cell transcriptomic efforts, we also need to pay attention to strategies acquiring consistent transcriptomic data for the high-resolution definition of human transcriptomes. Once distributions of transcripts in a limited number of cells are by and large defined, the identification of its expression and structural variations at single cell and single copy resolutions becomes easy.

It is obvious that a multi-national consortium is of immediate necessity, based on experiences from the International Human Genome Project. It will secure funding from governmental agencies and private sources, coordinate activities for different phases of the project, allocate tasks to its members, and organize efforts to improve technology and methodology. Other technical elements of the project are also worthy concerning. Is it necessary to separate transcriptomes of pathological nature from the physiological nature for cell-based effort assignments? Should we define non-coding RNAs, mRNAs, and small RNAs as separate efforts? What animal model systems should we use as control or validation tests? What new databases and algorithms should start to build? How do we incorporate parallel data from other omics-data, such as those of genetics, epigenomics, and proteomics?

The Human Transcriptomes Project is a larger project as compared to the Human Genome Project, where genome sequences are largely defined by population diversity and the number of sequence variations is finite in number and occurrence but transcriptomes are highly variable for each individuals and cells and influenced by genetic, epigenomic, and environmental factors. It is essential to have a common information platform to integrate all molecular data in a cellular context; it is equally important to have a standard protocol for all efforts to produce consistent transcriptomes. Nonetheless, it is not

***Corresponding author:** Jun Yu, CAS Key Laboratory of Genome Science and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, Tel: +86 10 8409 77; E-mail: junyu@big.ac.cn

**Received** September 28, 2015; **Accepted** October 01, 2015; **Published** October 05, 2015

**Citation:** Yu J (2015) The Human Transcriptomes Project: Is it hard? Next Generat Sequenc & Applic 2: e104. doi:10.4172/2469-9853.1000e104

necessarily an open-ended project. To sequence all transcripts of a cell is equivalent to an individual genome in tens of giga-bases. To work out all types and positions of RNA covalent modifications may be hard but it can be overarched to some specialists for detailed scrutiny.

It is time to start the Human Transcriptomes Project as genome sequencing is on the way to become a medical routine. Sooner or later, we will realize this; in order to not regret a lack of vision, let us act and act immediately.