Vol.2014-BIO-39 No.5
2014/9/19

Stem Cell Informatics Database: 多様なヒト細胞情報の統合化を目指したデータベースの構築

桜井都衣[1] 山根順子[1] 小林健太[1] 山根木康嗣[2] 谷口丈晃[3] 加藤有己[1] 藤渕航[1]

2006 年に iPS 細胞が発見されて以降，実際に幹細胞由来の人工組織が医療の現場で活用されるという時代の到来が現実味を帯びて来た．それと同時にこれまで以上に細胞というものを熟知し，個々の細胞の性質の判断基準を明らかにする必要性も著しく高まっている．

当研究室では，以前公開された SHOGoiN というヒト細胞解析データのためのデータベースの延長として，細胞というシステムを科学的な解析結果と知識に基づき明確に定義するためのデータベース（Stem Cell Informatics Database）の構築を行っている．現在の Stem Cell Informatics Database は，近年盛んに行われ始めているシングルセル技術を用いたトランスクリプトーム解析やメチローム解析データ及び実験条件などのメタデータ，ヒト組織画像とそこから抽出された細胞形態解析データ，そして解剖学的な位置情報に対応させたヒト細胞の分類表などを格納している．また今後は，細胞という複雑な概念のモデル化，語彙・知識の統合管理を可能にするためのオントロジーの構築や細胞情報解析のためのツールなども行い、搭載する予定である．

Stem Cell Informatics Database: a framework for a new repository on single cell assay data and diverse knowledge of human cells

Kunie Sakurai[1] Junko Yamane[1] Kenta Kobayashi[1] Koji Yamanegi[2] Takeaki Taniguchi[3] Yuki Kato[1], and Wataru Fujibuchi[†1]

Researchers have actively investigated potential applications of the induced pluripotent stem cell (iPS cell) for disease modeling, drug screening, and regenerative medicine since its discovery in 2006, and now it is about to start one of the projects for clinical trials. In parallel with the medical practice, there has been arising a need of more precise knowledge of the cell of its disposition, while facing a lack of the information system that enables us to store such complex biomedical knowledge regarding cells in well-organized way.

Here we introduce our cell knowledge repository called “Stem Cell Informatics Database” that is an extended work of the previous SHOGoiN database. It has been designed to integrate information comprehensively for defining cells with diverse knowledge and scientific data from biomedical research. In the Stem Cell Informatics Database, there are several indispensable contents, such as gene expression profiles and images of cells, curated assay metadata, and the cell taxonomy associated with anatomical location information. Stem Cell Informatics Database is now under development, and we are currently working on i) creating our own ontology to formally describe/model knowledge about the cells, and ii) developing analysis tools for gene expression data produced in single cell experiments. Depositing all of those in one database, this will provide a framework of integrative system for cell knowledge dictionary.

1. はじめに

近年の細胞工学研究の発展や iPS 細胞の臨床試験への応用にみられるように，細胞を人工的に作製し利用する時代が訪れようとしている．それと同時にその安全性の評価と品質管理のため，基準となる個々の細胞についての知識と品質管理の必要性が高まっている．現在は細胞レベルでのシステムの網羅的な解析（オミクス解析）技術が盛んに開発され,それに伴いビッグデータが産出される時代である．ビッグデータをいかに扱いやすい形式で効率的に保持するか，というのも大きな課題の一つである．

これまで当研究室では､SHOGoiN というヒト細胞解析データのためのデータベースを構築し，独自の細胞分類法を軸として，遺伝子発現データ，細胞画像データ，論文データ，そして細胞分化データを蓄積してきた．更に現在 Stem Cell Informatics Database はその延長として，シングルセル技術を基軸としたトランスクリプトーム解析やメチローム解析のデータを集積するなど，時代に合うデータベースへと発展している．また近い将来には，独自のオントロジーや，細胞を記述するための最小情報項目ガイドラインを国際協力の元に開発し格納することを予定している。これらを集約させ広く活用されることで，医学系研究の促進へと繋がることを期待している．

2. Stem Cell Informatics Database

こちらで報告する Stem Cell Informatics Database は，ヒト細胞に関する知識や解析データを格納するための統合データベースである（図 1）．現在は，細胞の遺伝子発現解析や形態学的な解析情報，解剖学的な局在情報を付加した細胞の分類表などを格納している．以下に Stem Cell Informatics

1 京都大学 iPS 細胞研究所
Center for iPS Cell Research and Application, Kyoto University
2 兵庫医科大学 病理学講座機能病理部門
Department of Clinical Pathology, Hyogo College of Medicine
3 三菱総合研究所 人間・生活研究本部
Advanced Business Division, Mitsubishi Research Institute, Inc.

Database における主な項目を挙げ，それらの詳細を説明する．

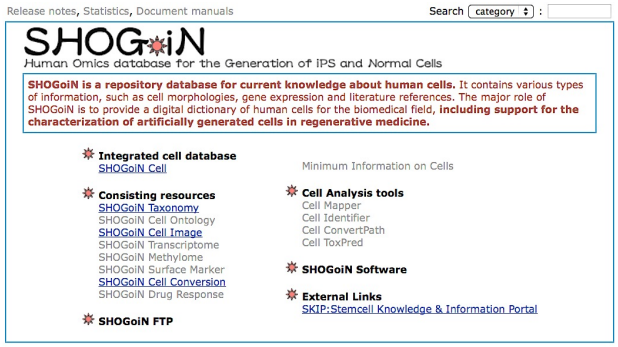

図 1　Stem Cell Informatics Database

2.1 細胞解析データ

以前公開された SHOGoiN データベースには、マイクロアレイなどの遺伝子発現解析データを 1063 件, また細胞画像データに関しては共同研究等で得られた組織切片画像と，そこから抽出した細胞の画像が計 638 件格納されている[1][2]．近年は更に技術の成熟により，次世代シークエンシングなどを用いたシングルセルレベルの遺伝子発現解析が盛んに行われている[3]．この方法により現在は，一種の組織に局在している細胞集団からの平均化された遺伝子発現情報ではなく，「個々の細胞」の遺伝子発現解析による更に詳細な細胞の分類が可能になっている[4][5]．これらのことからシングルセルを用いた遺伝子発現解析の重要性に注目し，新たに遺伝子発現解析データの収集を試みた．これらのデータに関しては，公共データベースのマニュアルキュレーションを行い，シングルセル技術を用いた実験データのみを収集し追加した．主な公共データベースとして，米国 NCBI の Sequence Read Archive [6]，Gene Expression Omnibus [7]や欧州 EBI の ArrayExpress [8]を用いた．その結果，現在 Stem Cell Informatics Database にはトランスクリプトーム解析データ 42 件, メチローム解析データ 7 件が蓄積されている．

更に，それらの実験のデザインには背景となる生物学的知識に加え，用いられた実験技術や材料に関する大量のメタデータにも有用な情報が含まれるため，それらの情報も統一された形式で保持している．

2.2 Cell Taxonomy sheet

従来の約 250 種類存在すると言われているヒトの細胞は，その形態や機能に関連して分類が行われてきた．しかし上述のシングルセルレベルの遺伝子発現解析を用いることによって，形態や機能の観察だけでは見つけることの出来なかった新規の細胞種が発見される可能性が考えられる．そのため既存の方法とは別に，新たな分類法・及び細胞の命名法を再考する必要性が生じた．そこで当研究室では，ヒトの体を構成する細胞の解剖学的な位置情報に基づいた分類を試みた．その結果，例えば以前まで体内の様々な部位で観察される繊維芽細胞，上皮細胞，平滑筋細胞などの細胞種を局在部位ごとに別の細胞として分類することが可能となった．分類された細胞は約 2050 種類にものぼり，更にこれらの細胞は解剖学的な分類を加味した ID を割り当てられ，体系的な管理が可能になると考えられる．

2.3 オントロジーと Minimum Information Standards

様々なリソースからの知識や情報の統合，共有，そして比較を行うためには，それら情報の形式や語彙が統一されることが重要であり，オントロジーはその役割を担うツールとして注目されている．近年，Seltmann らによって CELDA という細胞に関する知識を記述・モデル化するためのオントロジーが開発された[9]．しかし，CELDA は既存のオントロジーのクラスや語彙を統合する方法で開発されたため，クラスの重複や全体的な語彙の形式が統一されていないなど，いくつかの問題がみられる．そのため当研究室では，矛盾のない構造と十分な概念及び語彙を持つ独自のオントロジーの構築を試みている．また同時に，情報統合化を促進させるため重要性が注目されている，細胞情報を記述するための最小限の項目リスト（Minimum Information Standards）の作成も国際協力の下行っている．

3. おわりに

著しい研究技術の発展とともに，これまで当然とされていた知識や概念が覆されるということが起こりうる時代になっている．このことを踏まえると，新たな情報や知識の産生に対応しうる，柔軟な構造のデータベースを構築することが重要である．Stem Cell Informatics Database は現在開発中であり，既存の細胞の遺伝子発現データや画像データに加え，今後は更に多様なオミクス解析データの統合も視野に入れる必要がある．これらの網羅的な解析データや知識を統一された形式で体系的に集約し細胞に関連付けられれば，Stem Cell Informatics Database は細胞に関する辞書や新たな知識産生のためのツールとして共有され役立つことが期待される．

4. 参考文献

1) Hatano A et. al., Database: the journal of biological database and curation, bar046 (2011)
2) SHOGoiN, http://shogoindb.cira.kyoto-u.ac.jp/index.html
3) Tang F et al., Nat. Methods, 6, 377-382 (2009)
4) Jaitin DA et al., Science, 343, 776-779 (2014)
5) Trapnell C et al., Nat. Biotech., 32, 381-386 (2014)
6) Sequence Read Archive, http://www.ncbi.nlm.nih.gov/sra
7) Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/
8) ArrayExpress, http://www.ebi.ac.uk/arrayexpress/
9) Seltmann et al., BMC Bioinformatics, 14, 228 (2013)