美國癌症基因體圖譜計畫TCGA (The Cancer Genome Atlas) 簡介

鄒佩玲　吳昌俊

美國德克薩斯大學MD安德森癌症研究中心　基因體醫學部

摘　要

癌症基因體圖譜TCGA (The Cancer Genome Atlas)是一個以促進研究者對癌症的分子機制了解爲目標的實用資源。本篇文章將簡述這些資料的種類及性質，介紹標準分析報告，並摘錄包括神經膠質細胞瘤glioblastoma (GBM)，卵巢癌(ovarian cancer)等等及近數月來陸續新發表的重要研究結果，以期對從事相關研究的同仁們有所助益。

關鍵詞：癌症(Cancer)
基因體(Genome)
生物資訊(Bioinformatics)
資料庫(Database)

前言

癌症基因體圖譜TCGA(The Cancer Genome Atlas)[1]是由美國National Cancer Institute (NCI)及National Human Genome Research Institute (NHGRI)於2006年啓動的大型研究。此計畫目標是冀望透過更全面性，系統性的瞭解惡性腫瘤形成、生長、轉移等過程的分子生物基礎及病理機轉相關的基因體變化，以期促進癌症早期診斷及加速治療發展的腳步，更進一步而或能預防癌症的發生。

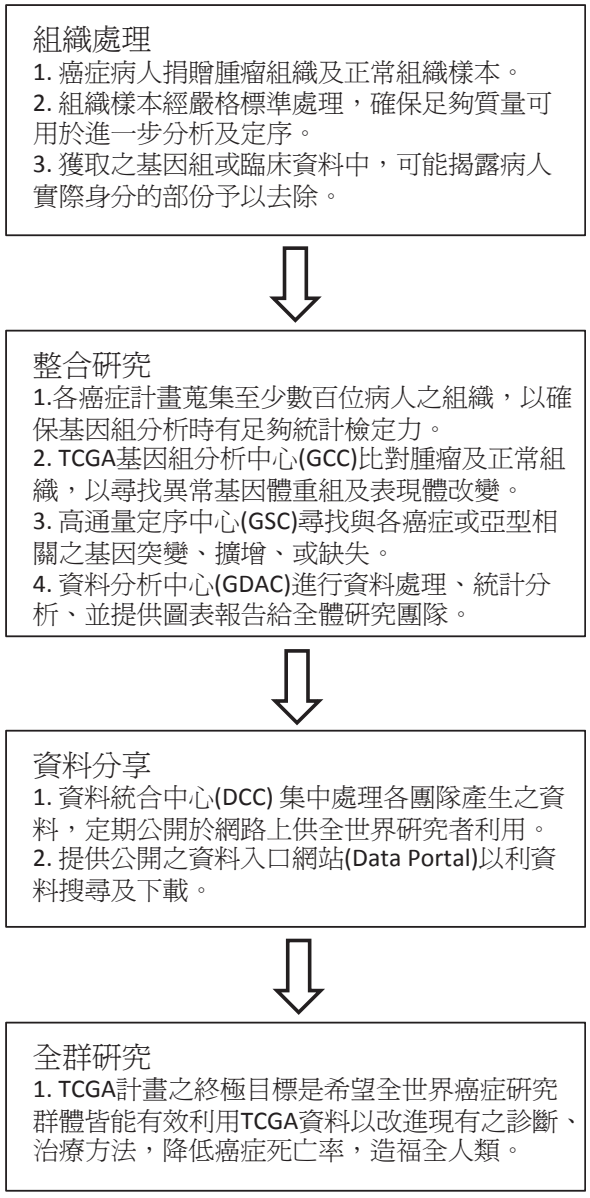
TCGA計畫的執行，乃藉由結合美、加數個大型醫學研究中心的組織樣本，儀器設備，及研究團隊，對特定癌症大規模地蒐集數百位病人的臨床病歷紀錄和腫瘤及其相對應正常組織樣本，血液樣本，進行全面的基因體資料擷取和整合性的分析，以期藉此促進對癌症的分子生物機制進一步的了解(資料處理及研究分工流程如圖一)。

此項計畫在神經膠質細胞瘤(Glioblastoma)[2]及卵巢癌[3]成功的證實了資源集中策略能有效加速癌症研究，近年已擴展到超過二十種腫瘤(表一)。截至2012年11月爲止，TCGA研究團隊已發表多篇論文於國際知名期刊中[2-12]。

最重要的是，此計畫的資料和大部份分析結果皆公開於網路上，可供瀏覽及下載。世界各地的研究者皆可共享此珍貴資源，並應用在自己的相關研究中。至今有七種癌症資料因相關論文已發表或接近發表(Glioblastoma, ovarian, breast, colorectal, endometrial, lung squamous

聯絡人：鄒佩玲　通訊處：Department of Genomic Medicine, Division of Cancer Biology, MD Amderson Cancer Center, University of texas, 1901 East Rd, 4SCR6.1314, Houston TX, 77030, USA

cancer, clear cell kidney cancer)，基於這些癌症資料之研究可不受限制發表於自己的論文中。另四種癌症(lung adenocarcinoma, papillary thyroid, head and neck cancer, skin melanoma)資料已公開，但基於此資料之論文須暫時受到限制，於2013年5月至7月陸續解禁後方能發表。其他十數種癌症資料因樣本數不足，解禁日期未定。詳細發表限制情況請參照 http://cancergenome.nih.gov/abouttcga/policies/publicationguidelines



圖一：TCGA團隊分工及資料分析流程圖。

本篇文章將簡單說明這些資料的種類及性質，介紹標準分析報告，及摘錄重要的研究發現，以期對從事相關研究的同仁們有所助益。

表一：美國癌症基因體計畫(The Cancer Genome Atlas, TCGA)包含之24種癌症及2012年11月止之病人數

癌症種類	病人數
Acute Myeloid Leukemia [LAML]	202
Bladder Urothelial Carcinoma [BLCA]	153
Brain Lower Grade Glioma [LGG]	222
Breast invasive carcinoma [BRCA]	919
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	122
Colon adenocarcinoma [COAD]	423
Glioblastoma multiforme [GBM]	600
Head and Neck squamous cell carcinoma [HNSC]	328
Kidney Chromophobe [KICH]	66
Kidney renal clear cell carcinoma [KIRC]	502
Kidney renal papillary cell carcinoma [KIRP]	117
Liver hepatocellular carcinoma [LIHC]	99
Lung adenocarcinoma [LUAD]	473
Lung squamous cell carcinoma [LUSC]	376
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC]	28
Ovarian serous cystadenocarcinoma [OV]	597
Pancreatic adenocarcinoma [PAAD]	57
Prostate adenocarcinoma [PRAD]	179
Rectum adenocarcinoma [READ]	169
Sarcoma [SARC]	29
Skin Cutaneous Melanoma [SKCM]	273
Stomach adenocarcinoma [STAD]	255
Thyroid carcinoma [THCA]	435
Uterine Corpus Endometrioid Carcinoma [UCEC]	512

一、癌症基因體圖譜計畫資料類別

癌症基因體圖譜計畫立眼於整合性的資料與系統性的分析。除了蒐集癌症病人臨床病歷資料外，對每個腫瘤樣本，亦取得Gene mutations (包括insertion/deletion), DNA copy number, mRNA expression, microRNA expression, protein expression, 及DNA methylation資料(詳見表二)。早期mRNA/microRNA transcriptome，及copy number由Affymetrix或Agilent microarray測量，自2011年起，更轉由新一代定序技術(Next Generation Sequencing)來獲得更快速精準之定性及定量的資料。

因爲可藉由病人編號將上述各類型之臨床及基因體資料配對，不論是統計癌症專一性基因體病變，或尋找癌症轉移或存活相關基因，甚至探索單一基因之作用機轉，或基因間的交互作用，都有很好的配對資料可供研究利用。除此之外，這類配對的臨床及基因體資料，也可應用於不同類型癌症之間分子生物機轉比較。

此外，某些癌症同時有對應(matched)或非對應(unmatched)樣本的正常組織mRNA/microRNA expression或methylation資料，此類型樣本所提供的訊息，更大有助於腫瘤發生(tumorigenesis)機轉方面的研究。近年來，TCGA更擴展資料的類別，加入運用reverse-phase protein array技術爲細胞中proteome及kinome定量資料，及運用定序技術獲得之染色體結構資料，可用於尋求癌症相關之染色體轉位(translocation)。

臨床及基因體資料可於TCGA資料門戶網站下載(TCGA data portal: http://tcga-data.nci.nih.gov/)

二、標準化基因體資料分析報告

爲了幫助臨床研究者及癌症生物學者更有效利用基因體計畫產生的資料，由位於波士頓的Broad Institute主導，聯合美、加數個生物資訊及計算機生物團隊組成基因體資料分析中心(Genome Data Analysis Center, GDAC)，設計標準化分析流程並定期將分析報告公開。內容包括MutSig Analysis：整合所有病人之基因變化，以統計分析尋找顯著好發突變之基因；GISTIC (Genomic Identification of Significant Targets In Cancer) Analysis：整合所有病人基因組之定量資料，以統計分析尋找染色體上好發擴增或缺失的局部區間，並列出此區間所含之基因；CNMF clustering：基於各種分子資料進行無監督(unsupervised)的non-negative matrix factorization clustering analysis，以尋找腫瘤可能的亞型；各種臨床相關及存活分析：尋找和癌症分期或存活等預後變項最相關之基因表現，突變，腫瘤亞型；基於分子途徑之分析(pathway analysis)；及其它整合各類型基因組資料的相關分析。這些報告不僅提供各式圖表以利生物學者建立或驗證研究假說，由於分析方

表二：TCGA資料類別

資料類別	簡介
Clinical	病人基本資料，治療進程，臨床分期(TNM)，腫瘤病理，存活情況等。
mRNA	由mRNA microarray, exon array 或 RNA-Sequencing測得之各基因mRNA表現量。
micro RNA	由microRNA microarray 或 microRNA-Sequencing測得之表現量。
Copy Number	由SNP microarray測得腫瘤對比於正常組織或白血球之染色體各區段(segment)比例。
Mutation	腫瘤組織定序資料對比於參考基因體序列得到之核苷酸變化(包含insertion, deletion)，及對應之氨基酸變化。
Protein	由reverse-phase protein array測量之約兩百種常見癌症相關蛋白質表現量，包括重要訊息傳遞(signal transduction)蛋白磷酸化程度。
Methylation	由methylation microarray測量之DNA甲基化程度。

法的標準化，統一化，大大地提昇了各資料間的可比較性。每月產生之分析報告可在分析中心網站下載：http://gdac.broadinstitute.org

三、已發表重要發現之摘錄

原則上TCGA研究團隊對每個癌症類別發表第一篇整合性研究報告，之後任何相關研究學者皆可自由發表運用此資料之論文。

第一篇Glioblastoma (GBM)研究報告於2008年發表於Nature[2]。整合基因突變和染色體局部擴增或缺失的結果，發現致癌機轉主要影響三個pathways: RTK/RAS/PI3K signalling (88%), TP53 signalling (87%), RB1/CDK4 pathway (78%)。後續研究發現GBM可分成四個亞型：Proneural, Neural, Classical, Mesenchymal[5]。各有特異之基因突變及不同局部染色體不穩定區域，其中Proneural亞型有較佳之存活。2010年另一篇論文同樣應用TCGA資料，發現此Proneural亞型之較佳存活，和DNA高度methylation有關，其中IDH1的突變占有重要的角色[4]。

對於卵巢癌(ovarian cancer)的研究於2011年發表[3]，首次運用新一代技術進行全基因組表現序列(exon)定序，發現突變好發於TP53, BRCA1, BRCA2, CSMD3, FAT3, NF1等基因。常見癌病變相關之pathways包括RB1/Cyclin/CDKN2A signaling (67%), BRCA/Homologous recombination repair pathway (51%), RAS/PI3K signaling (45%),及NOTCH signaling (22%)。 存活分析獲得了一組包含193個基因的預後指標，這些基因於細胞內mRNA表現量可預測卵巢癌病人之臨床預後。後續研究於2012年初發表在JAMA，發現具有BRCA1或BRCA2基因突變之侵襲性卵巢上皮癌(invasive epithelial ovarian cancer)病人與其他病人相比有較長之存活[7]。

研究團隊於2012年陸續於Nature期刊發表大腸直腸癌[9]，肺上皮細胞癌[10]，及乳腺癌[8]的研究報告。這些報告提供各癌症特徵之基因突變，染色體增殖缺失，及受影響之signaling pathways。乳腺癌研究發現，基於mRNA表現之腫瘤分組與2000年提出之乳腺癌亞型相呼應[13]，最惡性的basal亞型多爲triple negative (estrogen receptor/progesterone receptor/Her2 negative)，研究報告提出之基因突變及染色體不穩定區域爲標靶治療(targeting therapy)提供可行的目標[8]。肺上皮癌的研究發現7%的病人EGFR基因組擴增，這些病人可能對EGFR抑制劑治療有效[10]。大腸直腸癌報告證實最常突變基因包括APC, TP53, KRAS, PIK3CA等，並發現許多基因體變化與腫瘤的惡性程度有關[9]。

近兩年來已有超過 數十篇應用TCGA資料的論文發表，範圍涵蓋生物資訊，統計學，及癌症分子生物研究，如microRNA的分析[6]，DNA methylation分析[4]，及retrotransposition分析[12]。

結語

癌症基因體圖譜TCGA (The Cancer Genome Atlas)是一個以促進研究者對癌症的分子生物機制進一步的了解爲目標的珍貴資源。藉由蒐集並整合各類型臨床及分子資料，以及系統化，全面性的標準分析流程，個中蘊藏的豐富資訊，應可望成爲從事相關研究的同仁們的重要參考。不論是研究的初期利用TCGA資料由統計或生物資訊分析尋找實驗的方向，或是在實驗有重大成果時，將分子生物的發現與臨床預後相結合，皆可能有極大的幫助。雖然目前TCGA樣本來源多爲歐美人種，癌症致病機轉可能有地區性的差異，但在本土化大規模資料庫建立之前，仍能提供重要的資訊，並對未來的研究立下重要的基礎。

參考文獻

1. TCGA Home: https://wiki.nci.nih.gov/display/TCGA/TCGA+Home.
2. McLendon R, Friedman A, Bigner D, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008; 455: 1061-8.
3. Bell D, Berchuck A, Birrer M, et al. Integrated genomic analyses of ovarian carcinoma. Nature 2011; 474: 609-15.
4. Noushmehr H, Weisenberger DJ, Diefes K, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 2010; 17: 510-22.
5. Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010; 17: 98-110.

6. Genovese G, Ergun A, Shukla SA, et al. microRNA regulatory network inference identifies miR-34a as a novel regulator of TGF-beta signaling in glioblastoma. Cancer Discov 2012; 2: 736-49.
7. Bolton KL, Chenevix-Trench G, Goh C, et al. Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer. JAMA 2012; 307: 382-90.
8. Koboldt DC, Fulton RS, McLellan MD, et al. Comprehensive molecular portraits of human breast tumours. Nature 2012; 490: 61-70.
9. Muzny DM, Bainbridge MN, Chang K, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012; 487: 330-7.
10. Hammerman PS, Hayes DN, Wilkerson MD, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012; 489: 519-25.
11. Creighton CJ, Hernandez-Herrera A, Jacobsen A, et al. Integrated analyses of microRNAs demonstrate their widespread influence on gene expression in high-grade serous ovarian carcinoma. PLoS One 2012; 7: e34546.
12. Lee E, Iskow R, Yang L, et al. Landscape of somatic retrotransposition in human cancers. Science 2012; 337: 967-71.
13. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000; 406: 747-52.

Overview of the Cancer Genome Atlas Project

Pei-Ling Tsou, and Chang-Jiun Wu

Department of Genomic Medicine, Division of Cancer Biology, MD Anderson Cancer Center, University of Texas, Houston TX, USA

TCGA (The Cancer Genome Atlas), an integrated effort through the application of advanced genome technologies, is a treasure trove to facilitate our understanding of molecular basis of cancer. Here we introduce the collections of various cancer type and the standardized pipeline of data analysis, and summarize important findings from recent publications by TCGA teams.