EDITORIAL

Progress of cancer genomics

For hundreds of years, cancer has remained a mystery, threatening the life expectancy and quality of all human beings. Early studies of cancer concluded that all cancers are influenced by genetic variation (germline and/or somatic), environmental agents, and/or health behaviors. Thus, a comprehensive description of the human genome is essential in order to understand the biology of cancer.[1] Over the past decade, our knowledge of genomics has experienced substantial progress because of the rapid development of technology, which has dramatically advanced our studies of cancer.

Since the conclusion of the Human Genome Project (HGP) in 2003, the reference human genome sequence has provided the first comprehensive catalogue of the human genome and identified more than three million human genetic variants.[2,3] Despite the remarkable achievements following from the HGP, our knowledge of human genetic variation remains limited. Genetic association studies of the population provide a powerful approach to link genetic variants and diseases. To better identify disease-related variants, scientists launched the International HapMap Project alongside the 1000 Genomes Project.[4,5] These two projects have developed a detailed catalogue of genetic variants in the human genome, provide a blueprint for following studies, and launch us into a genome-wide association study (GWAS) era.

Association studies offer a classic strategy to study germline variants.[6] Integrating with prior knowledge of candidate genes or loci, we can identify heritable variants associated with cancers (development or survival status) in these regions.[7–12] However, a critical question remains; that is, not all reported associations can be robustly replicated in large studies or combined meta-analyses, possibly because of false positive, false negative, or population heterogeneity. With the application of high throughput genotyping technology, the GWAS has emerged as a more powerful and reliable tool for investigating the genetic architecture of complex diseases in a large sample size, without a prior hypothesis about a particular gene orlocus.[13] In the GWAS approach, several hundred thousand to millions of single nucleotide polymorphisms (SNPs) are assayed across the whole genome in a large sample size of thousands of individuals.[14] To date, GWA studies have led to the discovery of over 600 susceptibility loci of different kinds of cancers, including the loci that have been reported in Chinese populations.[15–21]

Nevertheless, two problems emerged and haunted post-GWA studies for many years. The first problem is missing heritability: only a small fraction of disease heritability could be interpreted by known susceptibility loci. Using prostate cancer as an example, 40 identified susceptibility loci only account for approximately 25% of the familial risk of disease. To deal with this problem, some studies turned to rare variants, structural variations, epistasis, and gene-environment interactions.[22–26] Compared with the common variants identified by GWAS (in general, minor allele frequency is more than 5%), rare variants are abundant in the human genome but are poorly detected by commonly used genotyping arrays, even after proper imputation. With the technology of exome sequencing, Thompson *et al.* successfully identified rare deleterious mutations in DNA repair genes as potential breast cancer susceptibility alleles.[22] Structure variants, including copy number variants (CNVs), inversions, translocations, microsatellite, repeat expansions, insertions of new sequence, complex rearrangements, and short insertions or deletions (indels) may also account for some of the unexplained heritability and are poorly captured by existing arrays.[27] Run of homozygosity (ROH) is a continuous or uninterrupted stretch of a genomic sequence without heterozygosity in the diploid state, which can be detected using GWAS data, but has been poorly investigated to date. Wang *et al.* explored the landscape and impact of ROHs on lung cancer.[23] Using an existing GWAS dataset including 1473 lung cancer cases and 1962 controls, a new region at 14q23.1 was identified to be consistently associated with lung cancer risk in the Chinese population, suggesting that ROHs may also be responsible for the unexplained familial risk of diseases. In addition, gene-gene or gene-environment interaction is thought to be one of the most important "dark matters" of missing heritability. However, it is still difficult to detect interaction in the current status of epidemiological study design, exposure assessment, and methods of analysis.[28] Some studies did some exploration, but interpretation of this statistic interaction is a great challenge.[24–26]

The second problem is the explanation of GWAS results: 88% of those variants from GWA studies fall outside of coding regions and have been difficult to interpret.[29] These problems have hampered our ability to pinpoint causal variants, identify genes affected by causal variants, and disentangle the mechanism by which genotype influences phenotype. Fortunately, the emergence of several large-scale genomic data sets generated by projects, such as the ENCyclopedia of DNA Elements (ENCODE), have revolutionized our ability to bestow potential function on GWAS identified variants. The ENCODE project is an international research consortium that aims to identify all functional elements in the human genome sequence.[30] It revealed that 80.4% of the human genome displays some functionality in at least one cell type. Integrating functional elements generated by ENCODE, Schaub *et al*. provided putative functional annotations for up to 80% of all previously reported associations.[31] Besides decoding new regulatory elements from human genomics, integrating data from multi-omics provides us new perspectives on GWAS results. On the basis of this strategy, Yao *et al*. identified 23 promoters and 28 enhancers potentially associated with colon cancer by using genomic and epigenomic information.[32] As the landscape of human transcriptome becomes more available, researchers attempt to establish a connection between genetic variants and gene expression, namely expression quantitative trait loci (eQTL). Many studies have demonstrated that GWAS signals are enriched with eQTL variants in a tissue-specific manner, highlighting their capability to help us understand the mechanisms underlying GWAS hits.[31,33]

Particularly, the potential for variants identified in GWA studies to predict the risk of complex diseases has been anticipated, but the usefulness of bringing these fundamental genetic findings to the bedside remains debatable.[34] Nevertheless, there are already a number of benefits of such genetic prediction over classical non-genetic models. For instance, genetic risk prediction is more stable over time than traditional risk factors, as a person's genetic sequence is absolutely constant throughout their life. Recently, Sun *et al*. reported that genetic score calculated by genetic variants discovered through an association study is an objective and better measurement of inherited risk of prostate cancer than family history.[35]

Another aspect of cancer genomic studies has focused on somatic alternations (e.g. mutations, CNVs, chromosome rearrangement). Unlike neutral germline variants, deleterious somatic mutation could act as a direct trigger of cancer, conferring oncogenic properties, such as growth advantage, tissue invasion and metastasis, angiogenesis, and evasion of apoptosis.[36] At the beginning of this century, studies on mutations were rare because of the complexity of the cancer genome and the limitations of technology. The emergence of massively parallel sequencing (MPS) revolutionized the entire enterprise. Since the first whole cancer genome sequencing by MPS in 2008, more than 10 000 cancer samples had been subjected to genome or exome sequencing by late 2013 in The Cancer Genome Atlas (TCGA) project (launched in 2005), let alone The International Cancer Genome Consortium (ICGC) project launched in 2009.[37–39] The explosion of genomic data quickly shed light on the mutational processes of cancer and revealed that cancer is much more complex than we originally thought; cancer mutation rates are much more variable, ranging from as low as one base substitution per exon (0.1/Mb) in some pediatric cancers to thousands of mutations per exome (~100/Mb) in certain mutagen-induced malignancies (such as lung cancer or melanoma); mutation patterns varied both across and within individual tumor types and some distinctive characteristics may reflect extrinsic factors like ultraviolet light or tobacco smoke, or intrinsic patterns such as DNA repair deficiencies.[40] To date, TCGA Research Network have published the genomic landscape of more than 10 types of cancers in top journals, identifying hundreds of potential "driver" alternations of cancers and classifying each cancer into a more detailed subtype by integrating mutli-omics data.[41–53] Such information could lead to more robust and personalized diagnostic and therapeutic strategies and provide a roadmap for developing new treatments.[54] However, our cancer genome catalogue is far from complete. Although a handful of cancer genes are found mutated at high frequency and could easily be detected, many more potential cancer-related genes are found mutated at much lower frequencies. As mentioned in a recently published paper, of 40 loci mutated at significant rates, 53% of the apparent driver mutations or focal copy number alternations were concentrated in six genes (*TP53*, *PIK3CA*, *ERBB2*, *FGFR1*/*ZNF703* and *GATA3*), and the remainder were dispersed across 34 genes.[55] Only eight of the genes were reported to mutate in at least 10% of breast cancers. It will be a great challenge to find these low frequency mutated genes with the current sample sizes. In contrast to point mutations in exons, our ability to discover and understand other types of driver alternations is still limited. As we cannot fully interpret the activation of cancer by mutations in known driver genes for each individual, many more important cancer drivers, including copy number alternations, chromosome rearrangements, and noncoding regions may hide in areas we cannot reach. With the development of sequencing technology and decreasing cost, we believe that we will deal with these problems and gather systematic information to inform a wider range of biological and clinical questions, and eventually realize personalized prevention, diagnosis, and treatment of cancer.

Disclosure

The author declares no conflict of interest.

Hongbing Shen[1,2]
[1]Department of Epidemiology and Biostatistics, School of Public Health and [2]Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China

References

1 Green ED, Guyer MS, National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011; **470**: 204–13.

2 International Human Genome Sequencing Consortium, Lander ES, Linton LM *et al*. Initial sequencing and analysis of

the human genome. (Published errata appear in *Nature* 2001; 412: 565; *Nature* 2001; 411: 720) *Nature* 2001; **409**: 860–921.

3 Hattori M. [Finishing the euchromatic sequence of the human genome.] *Tanpakushitsu Kakusan Koso* 2005; **50**: 162–8. (In Japanese.)

4 The International HapMap 3 Consortium, Atshuler DM, Gibbs RA *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–8.

5 1000 Genomes Project Consortium, Abecasis GR, Altshuler D *et al*. A map of human genome variation from population-scale sequencing. (Published erratum appears in *Nature* 2011; 473: 544) *Nature* 2010; **467**: 1061–73.

6 Shen H, Jin G. Human genome epidemiology, progress and future. *J Biomed Res* 2013; **27**: 167–9.

7 Zhang C, Li Z, Cao Q *et al*. Association of erythropoietin gene rs576236 polymorphism and risk of adrenal tumors in a Chinese population. *J Biomed Res* 2014; **28**: 456–61.

8 Liu Y, Zhang Q, Ren C *et al*. A germline variant N375S in MET and gastric cancer susceptibility in a Chinese population. *J Biomed Res* 2012; **26**: 315–18.

9 Cao S, Wang C, Huang X *et al*. Prognostic assessment of apoptotic gene polymorphisms in non-small cell lung cancer in Chinese. *J Biomed Res* 2013; **27**: 231–8.

10 Pan Y, Sun C, Huang M *et al*. A genetic variant in pseudogene E2F3P1 contributes to prognosis of hepatocellular carcinoma. *J Biomed Res* 2014; **28**: 194–200.

11 Cheng H, Deng Z, Wang Z, Zhang W, Su J. MTHFR C677T polymorphisms are associated with aberrant methylation of the IGF-2 gene in transitional cell carcinoma of the bladder. *J Biomed Res* 2012; **26**: 77–83.

12 Wang M, Chu H, Zhang Z, Wei Q. Molecular epidemiology of DNA repair gene polymorphisms and head and neck cancer. *J Biomed Res* 2013; **27**: 179–92.

13 Stadler ZK, Thom P, Robson ME *et al*. Genome-wide association studies of cancer. *J Clin Oncol* 2010; **28**: 4255–67.

14 Hardy J, Singleton A. Genomewide association studies and human disease. *N Engl J Med* 2009; **360**: 1759–68.

15 Wu C, Hu Z, He Z *et al*. Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nat Genet* 2011; **43**: 679–84.

16 Li S, Qian J, Yang Y *et al*. GWAS identifies novel susceptibility loci on 6p21.32 and 21q21.3 for hepatocellular carcinoma in chronic hepatitis B virus carriers. *PLoS Genet* 2012; **8** (7): e1002791.

17 Hu Z, Wu C, Shi Y *et al*. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet* 2011; **43**: 792–6.

18 Dong J, Hu Z, Wu C *et al*. Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat Genet* 2012; **44**: 895–9.

19 Dong J, Jin G, Wu C *et al*. Genome-wide association study identifies a novel susceptibility locus at 12q23.1 for lung squamous cell carcinoma in Han Chinese. *PLoS Genet* 2013; **9** (1): e1003190.

20 Shi Y, Hu Z, Wu C *et al*. A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. *Nat Genet* 2011; **43**: 1215–18.

21 Chen K, Ma H, Li L *et al*. Genome-wide association study identifies new susceptibility loci for epithelial ovarian cancer in Han Chinese women. *Nat Commun* 2014; **5**: 4682.

22 Thompson ER, Doyle MA, Ryland GL *et al*. Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles. *PLoS Genet* 2012; **8** (9): e1002894.

23 Wang C, Xu Z, Jin G *et al*. Genome-wide analysis of runs of homozygosity identifies new susceptibility regions of lung cancer in Han Chinese. *J Biomed Res* 2013; **27**: 208–14.

24 Wei WH, Hemani G, Haley CS. Detecting epistasis in human complex traits. *Nat Rev Genet* 2014; **15**: 722–33.

25 Chu M, Zhang R, Zhao Y *et al*. A genome-wide gene-gene interaction analysis identifies an epistatic gene pair for lung cancer susceptibility in Han Chinese. *Carcinogenesis* 2014; **35**: 572–7.

26 Zhang R, Chu M, Zhao Y *et al*. A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. *Carcinogenesis* 2014; **35**: 1528–35.

27 Manolio TA, Collins FS, Cox NJ *et al*. Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–53.

28 Chu H, Wang M, Zhang Z. Bladder cancer epidemiology and genetic susceptibility. *J Biomed Res* 2013; **27**: 170–8.

29 Hindorff LA, Sethupathy P, Junkins HA *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009; **106**: 9362–7.

30 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57–74.

31 Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res* 2012; **22**: 1748–59.

32 Yao L, Tak YG, Berman BP, Farnham PJ. Functional annotation of colon cancer risk SNPs. *Nat Commun* 2014; **5**: Article 5114.

33 Li L, Kabesch M, Bouzigon E *et al*. Using eQTL weights to improve power for genome-wide association studies: A genetic study of childhood asthma. *Front Genet* 2013; **4**: 103.

34 Pandey JP. Genomewide association studies and assessment of risk of disease. *N Engl J Med* 2010; **363**: 2076–7.

35 Sun J, Na R, Hsu FC *et al*. Genetic score is an objective and better measurement of inherited risk of prostate cancer than family history. *Eur Urol* 2013; **63**: 585–7.

36 Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science* 2013; **339**: 1546–58.

37 Ley TJ, Mardis ER, Ding L *et al*. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008; **456**: 66–72.

38 Cancer Genome Atlas Research Network, Wienstein JN, Collisson EA *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013; **45**: 1113–20.

39 International Cancer Genome Consortium, Hudson TJ, Anderson W *et al*. International network of cancer genome projects. *Nature* 2010; **464**: 993–8.

40 Garraway LA, Lander ES. Lessons from the cancer genome. *Cell* 2013; **153**: 17–37.

41 Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* 2014; **159**: 676–90.

42 Davis CF, Ricketts CJ, Wang M *et al*. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* 2014; **26**: 319–30.

43 Hoadley KA, Yau C, Wolf DM *et al*. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 2014; **158**: 929–44.

44 Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 2014; **513**: 202–9.

45 Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. (Published erratum appears in *Nature* 2014; 514: 262) *Nature* 2014; **511**: 543–50.

46 Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 2014; **507**: 315–22.

47 Brennan CW, Verhaak RG, McKenna A *et al*. The somatic genomic landscape of glioblastoma. (Published erratum appears in *Cell* 2014; 157: 753) *Cell* 2013; **155**: 462–77.

48 Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013; **499**: 43–9.

49 Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. (Published erratum appears in *Nature* 2013; 500: 242) *Nature* 2013; **497**: 67–73.

50 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61–70.

51 Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012; **489**: 519–25.

52 Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330–7.

53 Kandoth C, McLellan MD, Vandin F *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 2013; **502**: 333–9.

54 Sawyers C. Targeted cancer therapy. *Nature* 2004; **432**: 294–7.

55 Stephens PJ, Tarpey PS, Davies H *et al*. The landscape of cancer genes and mutational processes in breast cancer. *Nature* 2012; **486**: 400–4.