How statistics could inform breast cancer genetics research

Genes have been found to play an important part in the development of a number of diseases, including cancer. **Assistant Professor Audrey Fu**, a statistician working at the University of Idaho, has been investigating statistical methods for the causal inference of gene regulatory networks on diseases and her recent studies particularly focus on breast cancer. Her research hopes to explore potential differences in gene interactions between different breast cancer subtypes, which could improve the current understanding of cancer biology and help identify potential drug targets.

Cancer is one of the most devastating diseases in the world, causing a high number of deaths every year. Breast cancer is the most common cancer among women and is generally divided into three main subtypes, each one characterised by different patterns of gene expression. Understanding the genetic mechanisms behind these different subtypes could help further cancer biology knowledge, informing scientists on how different drugs might best address different genetic aspects of the disease.

THE STUDY OF DISEASE GENOMICS

The study of disease genomics is a complex and multi-faceted field of research. In recent years, multiple approaches have been used to investigate mechanisms behind gene regulation and their impact on diseases. Among these approaches, one entails studying gene regulatory networks and protein interaction networks, while another is comprised of genome-wide association studies that try to identify mutations and variants influencing particular diseases. Understanding how genetic variations (genotypes) influence specific diseases through gene regulatory networks is of huge interest to the scientific field, as it could inform research exploring specific forms of diseases and their treatments.

Over the course of her career, Dr Audrey Fu gained extensive experience in Bayesian statistics, a method of statistical inference that uses Bayes' theorem to update the probability for a specific hypothesis as more data becomes available. In recent studies, she has developed statistical models and computational strategies to help construct casual gene regulatory networks in diseases. Fu decided to focus her research on breast cancer, applying her method to its three main subtypes. This could help to reveal different mechanisms that may be behind each distinct breast cancer subtype.

BREAST CANCER SUBTYPES

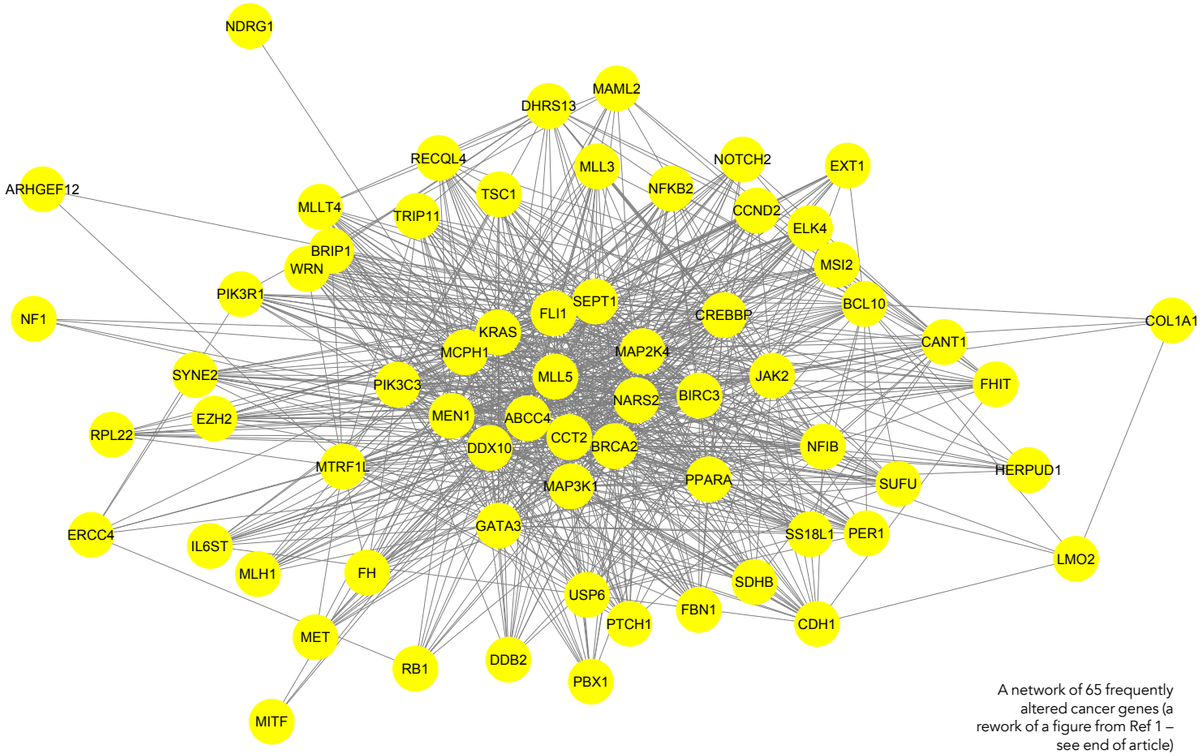
Breast cancer is a highly heterogeneous disease, meaning that it could be caused by a number of different genes and other factors, or by different interactions of these. There are three main subtypes of breast cancer: the oestrogen receptor positive (ER+) group, the HER2 amplified group, and the triple negative group. The ER+ group accounts for around 60% of breast cancers and is characterised by oestrogen receptors on the surface of cells that bind to oestrogen. The human epidermal growth factor receptor 2 (HER2) amplified subtype accounts for approximately 20–30% of breast cancers and those affected typically presents higher levels of the HER2 protein. Several effective targeted therapies have been developed for breast cancers falling in both the ER+ and HER2 amplified groups; these therapies target specific biological processes and can distinguish cancer and normal cells. Finally, around 10–20% of breast cancers are part of the triple negative group. A breast cancer is generally diagnosed as triple negative when oestrogen, progesterone and HER2 gene, the most common types of receptors known to prompt breast cancer growth, are not present in the cancer. This is the most aggressive breast cancer subtype and currently can only be treated with traditional therapies such as chemotherapy.

Fu attempts to answer a crucial question: what mechanisms drive the differences between breast cancer subtypes, especially when many genes may be involved?

Every one of these subtypes of breast cancer are characterised by a distinct pattern of gene expression and other particular molecular features. Researchers have also recognised additional, subtle classifications within these three groups. In her research, Dr Fu attempts to answer a crucial question related to these types of breast cancer: what mechanisms drive their differences, especially when many genes may be involved?

A STATISTICAL APPROACH TO BREAST CANCER GENOMICS

Past research has found that nuclear receptors (NRs), a class of proteins found within cells, play an important role in breast cancer. NRs, proteins responsible for sensing steroid and thyroid hormones, as well as other molecules, interact with each other to regulate the expression of their target genes. Fu's research aims to construct the gene regulatory networks involving NRs for the three different subtypes of breast cancer, in the hope of providing insight into the mechanisms that differentiate these groups. Her model integrates gene expression data, genotype data, and data on other binding transcriptional factors (TFs), proteins that control the rate of transcription of genetic information from the DNA to the RNA, binding it into a specific DNA sequence.



A network of 65 frequently altered cancer genes (a rework of a figure from Ref 1 – see end of article)

A SYSTEMS BIOLOGY APPROACH

Fu adopts a systems biology approach, a holistic way of trying to decipher complex biological systems based on the assumption that the whole picture is greater than the sum of its parts. Systems biology generally integrates different scientific disciplines, such as biology, computational and physical sciences.

In collaboration with Professor Kevin White at the University of Chicago, Dr Fu, a postdoc at the time, developed regression methods for analysing data collected by Dr Xiaoyue Wang, another postdoc in the lab, from large-scale RNAi knockdown assays for single and pairs of cancer genes (1). Dr Fu's statistical analysis enabled the construction of the first map of cancer gene interactions, which shows a high level of connectivity among these genes. Their analysis further demonstrated that these interacting genes are associated with survival time of cancer patients, again confirming the importance of accounting for gene interactions in the study of cancer.

Dr Fu now runs her own research group at the University of Idaho. Her team has recently developed a machine-learning algorithm that uses recent advances in graphical models to construct causal regulatory networks, taking into account both genetic variation and gene expression (2). This algorithm integrates different data sets collected from cancer patients, provided by The Cancer Genome Atlas (TCGA) consortium, to learn a network of genes and genetic variants. This is helping the team to hone in on genetic variants that result in changes in gene expression, and providing a picture of what genes may be regulating other genes. Fu plans to apply this method to the three breast cancer subtypes and study their different genetic mechanisms.

Dr Fu's work is unveiling differences in the complex genetic dynamics of different subtypes of breast cancer

USEFUL APPLICATIONS

As the performance of machine-learning algorithms improves, in future statistical methods will continue to play a huge role in studies that explore the causes and dynamics of complex diseases. Fu's work has the potential of contributing significantly to breast cancer research, as well as unveiling differences in the complex genetic dynamics of different subtypes of breast cancer. Her investigation might also offer insight into other areas to explore, such as key regulator genes or novel gene–gene interactions, ultimately leading to a better understanding of cancer biology, which would help to identify the areas that treatment drugs should target.

(1) Wang, X., Fu, A. Q., McNerney, M. E., & White, K. P. (2014). Widespread genetic epistasis among cancer genes. Nature communications, 5, 4828
(2) Badsha, M. B. & Fu, A. Q. (2017). Learning causal biological networks with generalized Mendelian randomization. biorxiv, https://doi.org/10.1101/171348.

Q&A

When and how did you first start trying to apply statistics to human genetics and diseases?

I worked with several fly biologists at the University of Cambridge during my first postdoc. I was greatly intrigued by the idea of using flies as an animal model to study human biology, and wondered whether and how we could study human biology directly. This curiosity brought me to my second postdoc at the University of Chicago, where I worked with Prof Kevin White and his lab members on genomics in cancer patients. This experience resulted in our 2014 publication in Nature Communications on an interaction map for cancer genes.

What do you think were your most important findings so far and why?

Since I am a statistician, 'findings' means useful methods that have wide applicability and that can help biologists gain insights into their experimental data. The most important 'finding' so far is the causal inference method my lab recently developed. This method can construct causal networks for genes and genetic variants using genomic data, with the directed edges in the network corresponding to causal (or equivalently, regulatory) relationships. I think that this method can help biologists better use existing large amounts of genotype and gene expression data. It also fixed several major issues in other software packages for causal inference.

How does your statistical approach differ from previous approaches to the investigation of cancer genomics?

My team is currently developing causal inference methods for genomics. Causal inference hasn't been used much in cancer genomics or disease genomics in general; existing analysis methods are mostly off-the-shelf ones that do clustering and regression. These methods typically use correlation, whereas similar correlation levels can arise from very different causal processes. With causal inference, we are able to integrate multiple types of genomic data and also use more information in the data. The result of applying this approach is a causal network of genes and genetic variants, learned directly from genomic data.

How could your research findings be applied to the study of breast cancer and/or within medical settings?

Our causal inference method integrates genotypes and gene expression, both of which are available through existing breast cancer consortia and are routinely collected in disease research. Our method constructs a causal network involving both the genetic variants and the genes. It suggests how a variant influences a number of genes through a network of genes. This helps us better understand the following questions: which genes play a more important role in breast cancer and what are their roles: a hub, or a super regulator, etc? Understanding these roles can be potentially helpful in designing effective treatments.

What are your plans for future investigation?

In the short-term, we are refining our causal network inference method, and extending it to account for more complex scenarios, such as when some nodes in the network are not measured. In the long run, we aim to extend the method to also incorporate other molecular phenotypes (such as DNA methylation, histone modifications) or clinical phenotypes (tumour size, metastasis or not).

We construct a causal network involving both the genetic variants and the genes. It suggests how a variant influences a number of genes through a network of genes

Detail

RESEARCH OBJECTIVES

Dr Fu is interested in developing statistical models and efficient computational methods for genomic data. Her research is focused on tackling problems related to human genetics and diseases.

FUNDING

National Institutes for Health (NIH, Pathway to Independence Award)

BIO

Dr Fu completed her PhD in Statistics (statistical genetics track) from the University of Washington, Seattle. She is Assistant Professor at the Department of Statistical Science, University of Idaho. Her research team develops statistical methods and machine learning algorithms for large quantities of biological data, such as data from the TCGA, 1000 Genomes, and GEUVADIS consortia.

CONTACT

Audrey Qiuyan Fu, PhD
Assistant Professor
Department of Statistical Science
Institute of Bioinformatics and Evolutionary Studies
Center for Modeling Complex Interactions
Department of Biological Sciences (affiliated)
University of Idaho
875 Perimeter Drive
MS 1104
Moscow
ID 83844-1104
USA

E: audreyf@uidaho.edu
T: +1 208 885 0132
W: http://webpages.uidaho.edu/audreyf/
@audreyqyfu
GitHub: @audreyqyfu