Social Science Research xxx (2016) 1–10

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Opportunities and challenges of big data for the social sciences: The case of genomic data

Hexuan Liu [a, c, d, *], Guang Guo [a, b, c]

[a] Department of Sociology, The University of North Carolina at Chapel Hill, USA
[b] Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, USA
[c] Carolina Population Center, The University of North Carolina at Chapel Hill, USA
[d] School of Criminal Justice, The University of Cincinnati, USA

ARTICLE INFO

Article history:
Received 30 July 2015
Received in revised form 8 April 2016
Accepted 13 April 2016
Available online xxx

Keywords:
Genomic data
Gene-environment interaction

ABSTRACT

In this paper, we draw attention to one unique and valuable source of big data, genomic data, by demonstrating the opportunities they provide to social scientists. We discuss different types of large-scale genomic data and recent advances in statistical methods and computational infrastructure used to address challenges in managing and analyzing such data. We highlight how these data and methods can be used to benefit social science research.

1. Introduction

The era of big data characterized by the explosion of information is upon us. For example, more than 300 million photos are uploaded to Facebook every day[1]; approximately 12 billion Google queries are conducted each month[2]; about two billion products were purchased from Amazon marketplace sellers in 2014.[3] Such big data not only generate economic value, but also promise new scientific discoveries. One of the best-known examples is the research of Ginsberg et al. (2009) that sought to predict influenza epidemics using 50 million Google search queries. This research led to the development of Google Flu Trends, a web tool that can detect flu outbreaks seven to ten days before they are reported by the Centers for Disease Control and Prevention (CDC).[4] As a more recent example, Preis et al. (2013) used online search data to predict "early warning signs" of changes in the stock market. Their test runs identified the financial crisis in 2008.

Social scientists have embraced enthusiastically the use of big data. In 2011, sociologists Sott Golder and Michael Macy published a paper in *Science* on changes in individuals' moods over time. Using 500 million public Twitter messages, they showed that people tend to have better moods when they wake up in the morning than later during the day, and the morning

* Corresponding author. Department of Sociology, University of North Carolina, Chapel Hill, NC 27500, USA.
E-mail address: hexuanl@live.unc.edu (H. Liu).

[1] "The Top 20 Valuable Facebook Statistics." (https://zephoria.com/social-media/top-15-valuable-facebook-statistics/).
[2] "By the Numbers: 60 amazing Google Search Statistics and Facts." (http://expandedramblings.com/index.php/by-the-numbers-a-gigantic-list-of-google-stats-and-facts/).
[3] "By the Numbers: 80 + amazing Amazon Statistics." (http://expandedramblings.com/index.php/amazon-statistics/2/).
[4] "Google Uses Searches to Track Flu's Spread." (http://msl1.mit.edu/furdlog/docs/nytimes/2008-11-11_nytimes_google_influenza.pdf accessed April 9, 2015).

http://dx.doi.org/10.1016/j.ssresearch.2016.04.016
0049-089X/

peak of positive moods is delayed by 2 h on weekends (Golder and Macy, 2011). At the 2014 annual meeting of the American Sociological Association, at least seven sessions were held focusing on big data.

In this article, we draw attention to one important type of big data—genomic data—and their implications for the social sciences. Advances in genomics are among the most spectacular scientific achievements over the past few decades. These advances have dramatically improved our understanding in a wide range of areas such as biology, health, medicine, and human nature. Over the past two decades, the development in genomic technology has resulted in a phenomenal reduction in the cost of genomic data collection. As an example, when the Human Genome Project was accomplished in 2003 (IHGSC, 2001; 2004), sequencing a single human genome for the first time cost $3 billion, and thousands of biologists and geneticists from the United States, the United Kingdom, Japan, France, Germany, and China spent 13 years on the project. By 2015, the cost of sequencing one individual's genome has dropped to less than $1,000, and the task can be done in hours.[5] Such genomic revolution has led to a paradigm shift from investigation of the functions of single genes to analyzing the role and structure of the whole genome in hundreds and thousands of individuals. DNA sequencing data represent only one type of genomic data. Other types of genomic data include epigenomic data and genome-wide gene expression data. These data, when combined with traditional social science data, could lead to advances in social science that were unimaginable even a decade ago.

This paper is organized as follows. We first outline opportunities that genomic data provide for the social sciences. We then introduce different types of genomic data that have been or will be available and their potential contributions to social science research. We devote the rest of the paper to introducing statistical methods for analyzing one type of genomic data that are best-known and have generated many valuable findings: the genome-wide genotype data.

2. Opportunities of genomic data for the social sciences

In the social sciences it is commonly assumed that human beings are homogeneous at birth and that differences across individuals are attributed to social, cultural, and environmental influences. This assumption has been challenged by rapid development in molecular genetics. In recent decades, considerable effort and resources have been devoted to discovering genetic causes of human diseases (Visscher et al., 2012). The National Institutes of Health (NIH)'s Catalog of Published Genome-Wide Association Studies (GWAS) of early 2015 includes more than 2000 publications that have established associations between thousands of genetic loci and human diseases as well as other traits (Hindorff et al.).

Reviewing evidence from behavior genetics based on biometrical analyses, Freese (2008) noted that most of social science outcomes at the individual level are genetically influenced to some extent, and that genetic effects on the outcomes must be mediated through a chain of biological and psychological mechanisms.

Many social science outcomes such as cognitive development, educational attainment, occupational status, binge drinking, and substance abuse are likely to be influenced by numerous interacting genetic and socioenvironmental factors. Incorporating genomic measures will help social scientists better understand the complex interplay among these socio-environmental and genetic factors. In the following, we outline specific ways social science research may benefit from incorporating genomic information (see Belsky and Israel, 2014; Belsky et al., 2013c; Boardman et al. 2012a, 2013, 2014, 2015; Conley et al. 2013a, 2015; Conley and Rauscher, 2013; Conley et al., 2013b; Domingue et al., 2015, 2014a, 2014b; Guo et al. 2008a, 2008b, 2015a, 2015b; Li et al., 2015; Liu and Guo, 2015; Liu et al., 2015; Mitchell et al., 2015, 2014, 2013; Perry Forthcoming; Pescosolido et al., 2008; Shanahan et al., 2008; Simons et al., 2011). Genomic sciences are still under rapid development and new types of genomic data have been produced all the time. Therefore, the ways in which social science research may benefit from genomic advances are likely to be extended considerably in the future.

First, studies of gene-environment interactions probably represent the most important opportunities for social scientists. Gene-environment interaction refers to the interdependence between an environmental effect and a genotypic effect. Gene-environment interaction implies that an environmental influence is sensitive to the effect of a genotype and vice versa. Ignoring gene-environment interactions forces us to estimate only an average genetic effect (averaged over all environments) or an average environmental effect (averaged over all genotypes), thus potentially dismissing genetic, environmental or both effects. For example, suppose we estimate a model in which body mass index (BMI) is predicted by variants in the *FTO* gene and educational attainment. Gene-environment interaction is present when the effect of *FTO* on BMI depends on education or when the effect of education depends on *FTO*. Frayling et al. (2007) reported that individuals who carry particular variants of the *FTO* gene were found to weigh, on average, 1.2 kg more than those who do not carry such variants. This effect of 1.2 kg is obtained without considering the environment. The effect may be smaller than 1.2 kg (or even absent) for some individuals under certain social conditions but greater for others.

Findings from gene-environment interaction studies can be used in the development of an intervention strategy if there is evidence for exogenous environmental influences (Conley, 2009; Fletcher and Conley, 2013; Guo et al., 2015b). The strategy removes or adjusts influences of social exposures resulted from genetic propensities (e.g., alcoholics cluster due to their shared genetic propensities to drinking). The strategy is based on the idea that genotypes are fixed, but social exposures might be alterable. A variety of gene-environment interaction models have been proposed, theoretically discussed and empirically

[5] "Ten years ago today, it was revealed that the human genome had been decoded. A medical revolution beckoned. So what happened next?" (http://www.independent.co.uk/news/science/ten-years-ago-today-it-was-revealed-that-the-human-genome-had-been-decoded-a-medical-revolution-beckoned-so-what-happened-next-2011016.html accessed April 7, 2015).

Please cite this article in press as: Liu, H., Guo, G., Opportunities and challenges of big data for the social sciences: The case of genomic data, Social Science Research (2016), http://dx.doi.org/10.1016/j.ssresearch.2016.04.016