BioData Mining

EDITORIAL Open Access

Mining beyond the exome

Davnah Urbach[1] and Jason H Moore[1,2,3*]

* Correspondence: jason.h.moore@dartmouth.edu
[1]Dartmouth College, Institute for Quantitative Biomedical Sciences, One Medical Center Dr., Lebanon, NH 03756, USA
Full list of author information is available at the end of the article

In the late 18th century, Erasmus Darwin, Charles Darwin's grandfather, advocated evolutionary theory as a mean to "unravel the theory of disease". More than 200 years later, although Darwinian medicine is regaining some ground after having been muzzled during the second half of the 20th century, genomics has largely outcompeted evolution and has acquired a dictatorial success as a tool for studying disease etiology [1]. From an evolution-inspired perspective, we have gradually drifted into the habit of focusing primarily on genomic data from sources such as genome-wide association studies (GWAS). As a result, understanding the how and why of human diseases and pathobiology has largely become a matter of crunching DNA sequences. Despite the popularity of GWAS, their reality remains unchanged: most of the susceptibility loci they allow to identify explain only a small fraction of the heritability of complex diseases [2]. A number of reasons for the so-called "missing heritability" have been proposed [2], and our goal is not to review them all. Here we primarily reiterate that there is more to discover than non-synonymous point mutations and suggest that amid genetic deserts and genetic islands, there is also more to explore than the coding regions of the genome. We then highlight the importance and the necessity of designing efficient methods to mine beyond the exome.

The premise of GWAS is the "common disease-common variant" hypothesis, which posits that common diseases are, at least partly, associated with DNA sequence variations or polymorphisms present in more than 1-5% of the population. It turns out that most allele frequencies battle to reach the 5% detection threshold of commercial genotyping arrays and the "common disease-rare variant" hypothesis is gradually taking precedence over its counterpart [2]. Hence, aiming for the rare variants using whole genome sequencing for example is one first step into the right direction [3]. A further step is to deliberately include synonymous polymorphisms among the genetic variants considered in association studies. Although largely disregarded, synonymous polymorphisms are about twice as numerous as non-synonymous ones [4] and are often found responsible for altered protein structure, function and expression level [5]. Accordingly, a considerable list of disease-associated synonymous polymorphisms is already available [5] and there are more to be found. Besides single nucleotide polymorphisms (SNPs), variation can also be structural: multi-kilobase genomic regions can be inserted or deleted (copy number variation, CNV), or they can be moved (copy neutral variation), within (inversion) or between (translocation) chromosomes [6,7]. Structural variants have already been shown to contribute to disease phenotypes [8,9], but with the help of high resolution GWAS purposely designed to detect them, there are undoubtedly more discoveries ahead [6,7].

Variants can adopt different forms but they can also occur in different locations throughout the genome. When given the choice between (quasi) random SNPs and

SNPs located in coding regions (gene-centric approach), choosing the latter is the safer bet [10]. However, the fact that more than 80% of the risk-associated variants identified so far fall outside of the coding regions suggests that there is a third option, namely the non-coding regions of the genome, including intergenic regions, introns and 3' and 5' untranscribed regions [11]. Non-coding regions harbor plenty of functional DNA, composed essentially of regulatory elements such as enhancers, promoters, insulators and silencer, and of non-coding functional RNA such as micro-RNA (miRNA). As the non-coding regions of the genome have gradually been revealing their secretes, evidence for their etiological importance has accumulated. Accordingly, genetic variation at regulatory elements [12-15] and at miRNA [16-18] has been found to play an important role in various diseases. Both better SNP coverage and whole genome sequencing will allow for a more methodological exploration of the non-coding regions of the genome.

There is more to the genome than we may have believed. Yet novel discoveries heavily rely on the availability of adequate and powerful analytical tools to exploit rich and complex data. In particular, progress in our understanding of the genetic architecture of common diseases requires efficient methods for merging different types of data and exploiting them simultaneously. Recent literature provides promising ideas on how to combine expert knowledge and crude genotyping data. Cowper et al. [14] for example suggest the use of genome-wide regulatory networks as a framework to incorporate biological knowledge to the analysis and interpretation of genotyping data, including data collected in the non-coding regions of the genome. This fits into a broader systems genetics approach to human disease [19]. Data are accumulating at a faster rate than methodological tools do. We suggest that there is room and urgent need for more ideas on how to analyze and integrate the different sources of information that we extract from both popular and remote regions of the genome. The last five years has focused on the task of manipulating large genomic data sets. Now is the time to integrate and synthesize these disparate sources of genomic information.

Author details
[1]Dartmouth College, Institute for Quantitative Biomedical Sciences, One Medical Center Dr., Lebanon, NH 03756, USA. [2]Dartmouth Medical School, Department of Genetics, One Medical Center Dr., Lebanon, NH 03756, USA. [3]Dartmouth Medical School, Department of Community and Family Medicine, One Medical Center Dr., Lebanon, NH 03756, USA.

Received: 11 May 2011 Accepted: 13 June 2011 Published: 13 June 2011

References

1. Gluckman PD, Low FM, Buklijas T, Hanson MA, Beedle AS: **How evolutionary principles improve the understanding of human health and disease.** *Evol Appl* 2011, **4**:249-263.
2. Manolio TA, *et al*: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**:747-753.
3. Holm H, *et al*: **A rare variant in *MYH6* is associated with high risk of sick sinus syndrome.** *Nat Genet* 2011, **43**:316-320.
4. Rish NJ: **Searching for genetic determinants in the new millennium.** *Nature* 2000, **405**:847-856.
5. Hunt R, Sauna ZE, Ambudkar SV, Gottesman IM, Kimchi-Sarfaty C: **Silent (synonymous) SNPs: should we care about them?** In *Single Nucleotide Polymorphisms: Methods and Protocols, Second Edition* Edited by: A KA 2009, 23-39.
6. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L: **Challenges and standards in integrating surveys of structural variation.** *Nat Genet* 2007, **39**:S7-S15.
7. McCarroll SA: **Extending genome-wide association studies to copy-number variation.** *Hum Mol Genet* 2008, **17**: R135-R142.
8. Craddock N, *et al*: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.** *Nature* 2010, **464**:713-U86.
9. Stankiewicz P, Lupski JR: **Structural variation in the human genome and its role in disease.** *Annu Rev Med* 2010, **61**:437-455.
10. Jorgenson E, Witte JS: **A gene-centric approach to genome-wide association studies.** *Nat Rev Genet* 2006, **7**:885-891.

11. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci USA* 2009, **106**:9362-9367.
12. De Gobbi M, *et al*: **A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter.** *Science* 2006, **312**:1215-1217.
13. Jia L, *et al*: **Functional enhancers at the gene-poor 8q24 cancer-linked locus.** *PLOS Genetics* 2009, **5**.
14. Cowper-Sal lari R, Cole MD, Karagas MR, Lupien M, Moore JH: **Layers of epistasis: genome-wide regulatory networks and network approaches to genome-wide association studies.** 2010.
15. Wright JB, Brown SJ, Cole MD: **Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated Single Nucleotide Polymorphism in colorectal cancer cells.** *Mol Cell Biol* 2010, **30**:1411-1420.
16. Calin GA, Croce CM: **MicroRNA signatures in human cancers.** *Nat Rev Cancer* 2006, **6**:857-866.
17. Esquela-Kerscher A, Slack FJ: **Oncomirs - microRNAs with a role in cancer.** *Nat Rev Cancer* 2006, **6**:259-269.
18. Wojcik SE, *et al*: **Non-codingRNA sequence variations in human chronic lymphocytic leukemia and colorectal cancer.** *Carcinogenesis* 2010, **31**:208-215.
19. Nadeau JH, Dudley AM: **Systems genetics.** *Science* 2011, **331**:1015-1016.

doi:10.1186/1756-0381-4-14
Cite this article as: Urbach and Moore: **Mining beyond the exome.** *BioData Mining* 2011 **4**:14.