Opinion

Diving Through the "-Omics": The Case for Deep Phenotyping and Systems Epidemiology

Robin Haring[1,2] and Henri Wallaschofski[2]

Abstract

Enabled by diverse high-throughput technologies, the rapidly evolving field of "-omics sciences" offers the potential to study health and disease in breadth and depth at the human population level. We have recently linked genomics and metabolomics to present the first genome-wide association study of metabolic traits in human urine providing new insights into the functional background of chronic kidney disease. We propose *systems epidemiology* as a novel approach to study the complexities of human pathophysiology by integrating various population-level omic-metrics and to identify new trans-omic biomarkers.

Introduction

Epidemiology involves the study of disease prevalence, incidence, and its risk factors. But facing the moderate effect size and the ever growing number of newly uncovered risk factors, the current risk factor era in epidemiology experiences increasing criticism (Fallin and Kao, 2011). Focusing mostly on a single risk factor related to a disease, the simplicity of this single-level paradigm has serious limitations. First, due to considerable interindividual differences in disease expression, these approaches only broadly predict who will have increased risk. Second, they do not consider feedback or feedforward effects such that changes to one risk factor change the effect of another risk factor. Third, current approaches are not well designed to evaluate complex interactions between multiple exposures and their dynamics encompassing human disease. Finally, phenotypes are often coded at different levels of detail and with different aims, further diluting statistical power and strengths of association.

Systems Epidemiology

Recent advances in high-throughput -omic platforms such as expression arrays and mass spectrometry, with their exquisite sensitivity, and specificity, have led to the possibility of accumulating a wealth of genetic, transcriptomic, proteomic, and metabolomic data to study health and disease in breadth and depth at the human population level. Based on the increased amount of detail available to describe an individual phenotype, we propose *systems epidemiology* as a new research field that integrates -omics together with physiological, epidemiological, and environmental data to create a systems network that can be used to predictively model multilevel causes of health and disease (Fig. 1). Further, the combination of complementary -omic levels could be implemented for the identification of novel trans-omic prognostic and diagnostic biomarkers. Therefore, we do think that deep phenotyping (Tracy, 2008)—the comprehensive and thorough description of the physical state of an individual—will be the corresponding principle of *systems epidemiology* on a population level, to lay the foundation for the analyses of dynamic feedback and interaction patterns among its multiple levels (Fig. 1).

Moreover, the here proposed *systems epidemiology* approach does not only concern the measurement of the molecular underpinnings of human disease, but also multiple environmental interaction components including behavioural, sociodemographic, and group levels that may influence health and disease. When considering human health in a wider perspective, it is clear that most major diseases are subject to environmental influences. For example, obesity has been recently reported to cluster in communities such that friends have an even more important effect on an individual's risk of obesity than genes do (Christakis and Fowler, 2007). Although epidemiology may be described as an effort in measurement, *systems epidemiology* could turn out as synthesis after measurement; an advanced pattern recognition approach integrating various -omics level (Fig. 1). This is not to say that epidemiology would relax its historic focus on populations, but need to absorb and apply the advancing scientific understanding at molecular and cellular levels to the study of health and disease in human populations.

If one accepts the case for deep phenotyping, including the limitation that increased costs will necessarily yield smaller sample sizes, there are at least three study design strategies

[1]Preventive Medicine & Epidemiology Section, Boston University School of Medicine, Boston, Massachusetts.
[2]Institute of Clinical Chemistry and Laboratory Medicine, University of Greifswald, Germany.

Systems Epidemiology
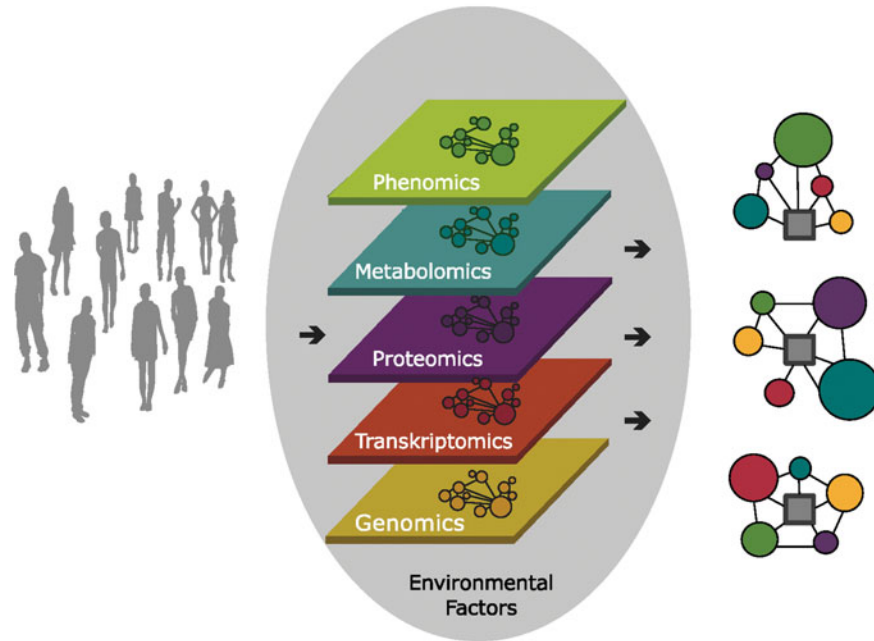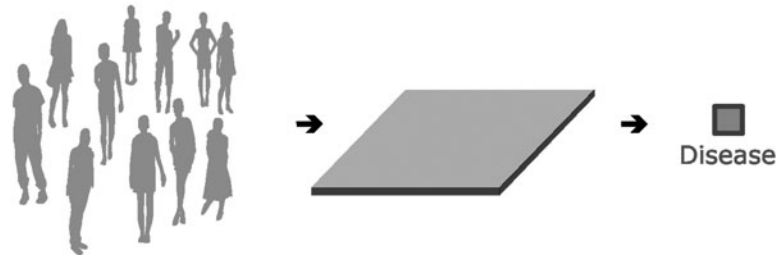

Single-level Epidemiology

FIG 1. Systems epidemiology versus the classic single-level paradigm to study health and disease at the human population level. Integrating various population-level omic-metrics including the **Phenome** (physical traits such as body height, weight, or specific personality characteristics), **Metabolome** (complete set of small-molecule metabolites to be found within a biological sample), **Proteome** (entire set of proteins expressed by a genome, cell, tissue, or organism), **Transcriptome** (information about the expression of individual genes at the messenger ribonucleic acid level), **Genome** (complete set of genes in the human organism), and **environmental factors** (behavioral, sociodemographic, and group levels), as well as the complexities of its interactions will be critical for developing the most effective diagnostic techniques in systems epidemiology. The understanding of each system-level component is also crucial in understanding the pathophysiology of human disease (gray squares), here shown as a function of subnetworks of a complex multiomics network (each coloured node in the subnetwork represents an -omic level, whereas node sizes are proportional to the strength of disease association and *links* between nodes indicate trans-omic relationships). The systems epidemiology approach is contrasted by the simplicity of the single-level paradigm in classical epidemiology focusing mostly on a single risk factor or omic-level related to a disease, respectively.

that might be employed in *systems epidemiology*. The first involves a longitudinal study design making multiple measurements over time with fairly short time periods between measurements. Data, for example obtained yearly over 10 years, would enable the investigator to closely monitor subclinical disease progression and to detect dynamic changes in the nature of the phenotype over time. A second operational approach accounts for the fact that for many biomarkers the within-subject variability is larger compared to the change in the biomarker over time. Therefore, a mean biomarker value calculated based on two or three blood draws spread over the day are likely to eliminate this within-subject variability. Finally, the case of deep phenotyping is likely to reduce the inaccuracy and misclassification of disease outcomes present in most epidemiological and clinical studies, by increasing an individual's phenotypic information and refining risk classification.

Network analytic methods provide the computational framework for data integration and biomarker selection in *systems epidemiology* (Adourian et al., 2008). For example, network-based computational approaches and longitudinal

data from the Framingham Heart Study revealed a number of surprising insights into the dynamics of smoking (Christakis and Fowler, 2008), development of obesity (Christakis and Fowler, 2007), and metabolic determinants of diabetes risk (Wang et al., 2011). Similarly, network-based analyses of known disease-gene associations revealed a number of surprising connections between diseases, forcing us to rethink apparently distinct pathophenotypes and their nomenclature (Goh et al., 2007). Taken together, the proposed integration of population-level omic-metrics promises to advance our understanding of human disease (Barabasi, 2007; Barabasi, et al., 2011; Loscalzo et al., 2007) and to enable the identification of new trans-omic biomarkers (Rantalainen et al., 2006).

We recently provided a proof of concept for the integration of population-level omic-metrics employed in *systems epidemiology* using genomic and metabolomic data for the first genome-wide association study of metabolic traits in human urine (Suhre et al., 2011). Through the identification of genetic variants related to metabolism, specific "genetically determined metabotypes" have the potential to uncover additional risk factors for common diseases and may provide new insights into the pathophysiology of these diseases. Using nuclear magnetic resonance spectroscopy to quantify 59 metabolites in urine from 862 male participants of the population-based epidemiological Study of Health in Pomerania (SHIP), we identified genetic variants that have been previously linked to important clinical outcomes including chronic kidney disease and coronary artery disease. The revealed plausible relationships between the associating metabolic traits and the genetic variants' encoded protein functions provided new insights into the metabolic basis and functional background of related pathophysiological processes. Thus, the study of genotype-dependent metabolic phenotypes may provide new functional insights for many disease-related associations and may constitute potential trans-omic biomarkers for diagnosis and monitoring (Suhre et al., 2011).

Challenges

Although disease manifestations have been shown to be derived via different pathways in different individuals, epidemiological research using multiple phenotypic levels to study human disease is scarce. Especially results from genome-wide association studies, perhaps the most vibrant research field of the last 5 years, have thus far proven to be surprisingly disappointing, partly because of the unexpected complexity of the human genome and the difficulties in accurately and unequivocally describing human phenotypes (Maher, 2008). Furthermore, biological systems exhibit robustness and dynamic stability where phenotypic changes are fairly resistant to scattered omic-level fluctuations (Hillenmeyer et al., 2008). Thus, competing risks are buffered by regulatory networks that use alternative mechanisms to ensure phenotypic stability (Nobrega et al., 2004). For example, competing risk patterns emerging from the nearly 600 genome-wide association studies reported nearly 800 significant single nucleotide polymorphisms (SNP)-trait associations with few variants having large effects, but most having small effects (Manolio, 2010).

Another important challenge may be the lack of biomarker standardization and harmonization. As advanced computational modeling techniques are used, well-integrated platforms and data sets are required to conquer the intersection of

highly dynamic parameters and to assess the relationships between omic-level metrics and their phenotypic manifestations (Connor et al., 2010). Furthermore, the proposed study design enables a dynamic exposure assessment requiring a close iteration between experimental data input and theoretical modelling (Kohl et al., 2010). For instance, it is increasingly recognized that the understanding of complex metabolic and cardiovascular diseases or cancer requires an integrated analyses of its molecular and cellular components, as well as their relationships, pathways, and interconnectivity. To address the risk of increased data noise and false positive findings in this extremely data-rich research environment resulting from high-throughput multiomics technology, the suggested application of high-performance computing technologies may facilitate high-volume data analysis and close the scientific gap between increasing amount of data, correlation identification, and plausible causal pathways (An, 2010).

Conclusions

The network-based integration of deep phenotypes in *systems epidemiology* may provide a novel approach to study the complexities of human pathophysiology, and to account for the paradox of ever-increasing measurement capabilities followed by decreasing abilities to translate basic mechanistic knowledge into clinically effective therapeutics (An, 2010; Lenfant, 2003). Thus, *systems epidemiology* aims to account for the large variability of interindividual disease onset, manifestation, and progression and may thereby support an individualized medicine grounded on a multiscale and non-reductionist analytical approach.

Acknowledgments

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

Adourian, A., Jennings, E., Balasubramanian, R., Hines, W. M., Damian, D., Plasterer, T. N., et al. (2008). Correlation network analysis for data integration and biomarker selection. Mol Biosyst 4, 249–259.

An, G. (2010). Closing the scientific loop: bridging correlation and causality in the petaflop age. Sci Transl Med 2, 41ps34.

Barabasi, A. L. (2007). Network medicine—from obesity to the "Diseasome." N Engl J Med 357, 404–407.

Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet 12, 56–68.

Christakis, N. A., and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. N Engl J Med 357, 370–379.

Christakis, N. A., and Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. N Engl J Med 358, 2249–2258.

Connor, S. C., Hansen, M. K., Corner, A., Smith, R. F., and Ryan, T. E. (2010). Integration of metabolomics and transcriptomics

data to aid biomarker discovery in type 2 diabetes. Mol Bio-syst 6, 909–921.

Fallin, M. D., and Kao, W. H. (2011). Is "X"-was the future for all of epidemiology? Epidemiology 22, 457–459; discussion 467–458.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A. L. (2007). The human disease network. Proc Natl Acad Sci USA 104, 8685–8690.

Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., et al. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science 320, 362–365.

Kohl, P., Crampin, E. J., Quinn, T. A., and Noble, D. (2010). Systems biology: an approach. Clin Pharmacol Ther 88, 25–33.

Lenfant, C. (2003). Shattuck lecture—clinical research to clinical practice—lost in translation? N Engl J Med 349, 868–874.

Loscalzo, J., Kohane, I., and Barabasi, A. L. (2007). Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Mol Syst Biol 3, 124.

Maher, B. (2008). Personal genomes: the case of the missing heritability. Nature 456, 18–21.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. N Engl J Med 363, 166–176.

Nobrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V., and Rubin, E. M. (2004). Megabase deletions of gene deserts result in viable mice. Nature 431, 988–993.

Rantalainen, M., Cloarec, O., Beckonert, O., Wilson, I. D., Jackson, D., Tonge, R., et al. (2006). Statistically integrated metabonomic-proteomic studies on a human prostate cancer xenograft model in mice. J Proteome Res 5, 2642–2655.

Suhre, K., Shin, S. Y., Petersen, A. K., Mohney, R. P., Meredith, D., Wagele, B., et al. (2011). Human metabolic individuality in biomedical and pharmaceutical research. Nature 477, 54–60.

Suhre, K., Wallaschofski, H., Raffler, J., Friedrich, N., Haring, R., Michael, K., et al. (2011). A genome-wide association study of metabolic traits in human urine. Nat Genet 43, 565–569.

Tracy, R. P. (2008). "Deep phenotyping": characterizing populations in the era of genomics and systems biology. Curr Opin Lipidol 19, 151–157.

Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., et al. (2011). Metabolite profiles and the risk of developing diabetes. Nat Med 17, 448–453.

Address correspondence to:
Robin Haring, Ph.D.
Institute of Clinical Chemistry and Laboratory Medicine
University of Greifswald
Ferdinand-Sauerbruch Strasse
17475 Greifswald, Germany

E-mail: robin.haring@uni-greifswald.de