Fenger, J Genet Disor Genet Rep 2012, 1:1
http://dx.doi.org/10.4172/2327-5790.1000e104

Journal of Genetic Disorders & Genetic Reports

Editorial

A SCITECHNOL JOURNAL

Genetic Networks in Heterogeneous Populations

Mogen Fenger[1*]

The core in biological organisms is the information harbored in the genome, which encode the entire blue-print of functionality and regulation of processes in cells and integration of multicellular organisms. Population genetics aims at identifying genes of importance for biochemical and physiological pathways particularly with scope of revealing genetic causes of diseases. Although the rapidly growth in technology we have experienced recently has given us unprecedented opportunities to perform genetic studies, all the promises may not have been fulfilled entirely as hoped. The reasons are several, but it seems that two issues have been partially neglected, namely that study populations are all heterogenous and genes are not solitary but function in networks. This is not because these issues are not acknowledged, but disentangling the complexity of the genome and the encoded physiological processes are a staggering endeavor. Not the least the computational burden involved has set the limit of many approaches.

Imaging a diploid organism like *Homo sapiens* with 23 sets of chromosomes only harboring one mutation in each chromosome. In the meiotic process the number random gametes will amount to approximately 8.4 million. Imagine further that all gametes are viable then the number of possible zygotes amounts to more than $7*10^{13}$ or more than 11.000 fold the number human beings living on planet Earth today. Probably many of the gametes or zygotes are not viable, but then again single nucleotide polymorphisms and mutations discovered at present runs in the millions, and the number will increase as the exonic and full-genome sequencing projects proceed. Thus, no two human beings will ever be genetically identical including monozygotic twins as they may differ epigenetically.

Consider now physiological processes like glucose/fat metabolism or blood pressure regulation which are regulated by say 100 genes in an integrated network. Again, if just one mutation is present in each gene then the ensemble of networks with exactly the same topology that can be constructed is 10^{30}. This would map to as many physiological states and dynamics, which are impossible to study in practice. Adding to the number of genes, their alternative spliced forms, the vast number of posttranslational modifications of proteins, non- protein regulatory elements (metabolites, small regulatory RNAs), epigenetic modifications, non-genic regulatory and genome-organizing structures, and not the least interactions and communications between cells in a multicellular organisms like humans, the combinatorial space of interactions and hence phenotypes is (almost) infinite.

The members of the ensemble of genetic networks can be mapped as a "continuum" reflecting the physiological states they define (Figure 1). Neighboring networks are genetically distinct by variations in one or several genes or non-genetic regulatory structures but may appear physiological similar, suggesting that most genetic variations have small effects. Genetic variations may also be balanced in such a way that the physiological states appear similar and in practice are indistinguishable. The sensitivity to factors outside the network is encoded in the genome, and it is the variations in the process-specific genes and regulatory structures that determine the range of the response to an external perturbation. Thus, identifying genetic network is not simple and transparent: functional networks are multipartite structures, which in addition are not necessary secluded entities but rather interact with other networks (e.g. the glucose and the fat metabolism are highly intermingled processes). Nevertheless, it may be possible to define a reasonable number of sub-ensembles of networks to be interpretable (as indicated in the Figure), and hopefully even so that the genes of most importance in the sub-ensembles may be few (even "hub-like") to make disease prevention and treatment feasible and increasingly personalized.

Many strategies have been implemented to identify genes influencing a trait or causing a disease but for any approach to be successful two fundamental issues has to be addressed: any study population is physiological heterogeneous caused by the vast genetic variations present, and genes and non-genic genetic factors do always convey their information in networks i.e. by interactions.

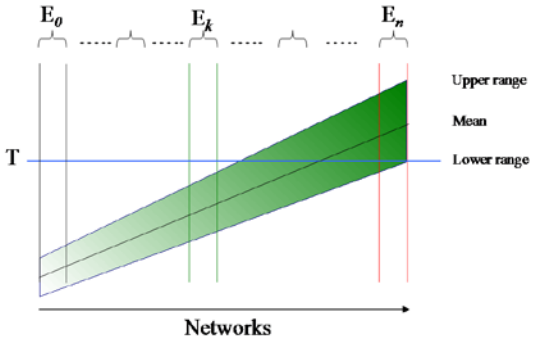

Figure 1: The ensemble of networks with exactly the same topology but differences in genetic variations is order according to increasing values of a trait, e.g. diastolic blood pressure. The entire ensemble is partitioned into an a *priori* unknown number of sub-ensembles E*i* (indicated by the sets of two vertical lines) by the LCA-SEM procedure. The networks in each sub-ensemble may arise by successively added mutations and/or be balanced networks with mutations in different parts of the network with opposing effects. "Mean" indicates the mean of the trait at the time of measurement, while the ranges indicates the limit of the trait values conditional on the genotype in the network. The darker grading indicates an increased propensity to develop a clinical endpoint e.g. diastolic hypertension indicated by the threshold line T.

***Corresponding author:** Mogen Fenger, Copenhagen University Hospital at Hvidovre, Department of Clinical Biochemistry, Genetics, and Molecular Biology, Kettegaard All 26, 2650 Hvidovre, Denmark, E-mail:mogens.fenger@hvh.regionh.dk

Received: July 26, 2012 **Accepted:** July 27, 2012 **Published:** July 29, 2012

doi:http://dx.doi.org/10.4172/2327-5790.1000e104

Population Heterogeneity

Population heterogeneity refers to the mixture of otherwise homogeneous subpopulations. In its extreme every subject defines a homogeneous subpopulation taken in consideration that every individual harbours a unique genome wide genotype [1-3] and no further analysis of heterogeneity is needed. This is however not fruitful as we will gain no knowledge of the genetic structures that are common to all subjects in a species. Rather, we should look for similarities that cluster subjects into more physiological homogeneous subpopulations. Importantly, focus should be on defining physiological states as these are defined by variations in the same genetic networks in all subjects of a species. Collapsing physiological variables into a dichotomous (or polytomous) variables e.g. affected/not affected will lose information [4], as disease and health merely are descriptors on an almost continuous functional scale of the same physiological process for which some physiological states are defined as disease. Hence, proceeding with the prevailing dichotomy of affected/not affected will for statistical and most certainly for biological reasons generally be of limited value.

The application of appropriate cluster algorithms to identify homogeneous subpopulations is generally ill posed, as no universal formal criteria for the best clustering are available. This problem is accentuated by the remoteness of the measured variables or features usually available in population genetics. Many of the well-known classification procedures implement some data-reduction e.g. dichotomizing continues data or focusing on subset of variables [5], but caution should be observed as any manipulation and reduction of the data space are prone to lose information as mentioned above.

Allocating subjects to subpopulations falls in the area of modeling hidden or latent variables as the number of subpopulations are not known *a priori*. This can resolved by applying the concepts of latent class (LCA) or latent profile analysis (LPA) in a structural equation framework (SEM) [1,2,6-8]. The philosophy of the LCA/LPA-SEM approach is to model physiological processes, and not particular outcomes like hypertension and therefore the most appropriate study population would simply be a random ascertainment of subjects as all subjects provide information of the physiological processes. Genetic structures and variations are not necessarily modeled directly, but are embedded in the SEM structure and are mapped or reflected by the measured manifest variables. Modeling in this framework addresses two pivotal issues in complex data: resolving the heterogeneity in the population, and simultaneously evaluating the data structure within the sub-populations [1,2,9]. This approach outperforms most other classifications methods in almost all aspects [10] and embraces the so-called genetic admixed population approach [11].

An emerging line of methods of particular interest is approaches using ensembles of classification functions [12]. Ensemble techniques combines several objective functions or decision algorithm (classifiers) to solve the same task, that is classification or clustering of data *in casu* subjects into homogenous subpopulations [13,14]. The approaches are attractive when features in multi-source or distributed data sets are completely disjoint or only partially overlaps, or if access is limited to a subset of objects in a data set. Thus, the problem of missing data and hence decreased power may be circumvented to some extent, when data from several sources are combined, and represent a potential alternative to imputing e.g. missing genetic data.

Networks in Biology

Graphs are visual and mathematical descriptions of complex systems and networks [15,16]. There is no exact universal definition of complex systems, but informally a complex system is a large network consisting of simple components without a central control in which complex behavior or mechanistic emerges. Thus, complex systems seem to be self-organized, self- supporting or "floating". The word "simple" has not to be taken too literally as simple components may be complex by itself, e.g. a whole cell. Complex systems are naturally non- linear, as the behavior or outcome of the system is governed by complex interactions rather than just the sum of its components activity. A graph is a collection of nodes (genes) connected by links. In its basic form the links are Boolean (0/1), but the weights as well of the types of the links can be included, for instance interaction (epistasis) between genes [1,2]. In addition, node characteristics (e.g. genetic variance or effects) can be incorporated [17].

Several measures have been developed to characterize the topology of a network [18] and is essentially related to the flow of information in the network (signal transduction, metabolic paths etc.) [19]. The issue is complex, but the important point is that the activity of a network is not necessarily defined by the shortest path defined by the Boolean adjacent matrix, but rather by the information carried. Many networks posses' two essential properties: scale-free degree distribution and small-world property. Scale-free degree distribution refers to the power-law distributions of the connections between nodes. The small-world property is characterized by a very slow increase in average path length between nodes with the size of the network. However, the relevance of these concepts in biology has recently been disputed [20], as the structure and dynamics of biological networks often differs from physical, social, and communication networks from where the concepts has been developed [21,22]. Although network theory definitely is pivotal to biological systems, transformation of theories and conclusions from one area to another may not be straightforward.

Conclusions

It is widely recognized that traits are heterogenous and that biological processes are the result of interacting genes, proteins and metabolites of any kind embedded in complex functional networks. The traditional genetic approach i.e. that single variations are causative is very limited, as the vast majority of genetic variations do not have any main effect, but their importance emerge in the context of interactions in networks [2]. The complexity of biological systems is staggering and understanding and integrating the wealth of particular genetic data in a medical context requires new approaches and techniques. Fortunately many new approaches are emerging increasingly embracing the nature of biological systems, particularly the recent developments in network theory including the concept of modularity [23], complex theories as stochastic block modeling [24], statistical mechanics [25,26], and information theory [27].

The focus is on defining community structures on any scale. Within biology networks this translates into defining sub-structures in a genetic network of particular importance for the physiological process, which could be imagined to differ for same process in different subpopulations [2]. Still in the developmental state, these recent approaches however seem promising in elucidating an important medical issue: resolving the genetics of diseases.

doi:http://dx.doi.org/10.4172/2327-5790.1000e104

References

1. Fenger M, Linneberg A, Werge T, Jorgensen T (2008) Analysis of heterogeneity and epistasis in physiological mixed populations by combined structural equation modelling and latent class analysis. BMC Genet 9: 43.
2. Fenger M, Linneberg A, Jorgensen T, Madsbad S, Sobye K, et al. (2011) Genetics of the ceramide/sphingosine-1-phosphate rheostat in blood pressure regulation and hypertension. BMC Genet 12: 44.
3. Fenger M (2008) Modelling genetic and physiological heterogeneity. In Genetic inheritance patterns. Nova Science Publisher.
4. Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25: 127-141.
5. Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23: 2507-2517.
6. Skrondal A, Rabe-Hesketh S (2007) Latent variable modelling: a survey. Scand J Stat 34: 712-745.
7. Bollen KA (1989) Structural Equations with Latent Variables. John Wiley and Sons.
8. Muthen BO (2002) Beyond SEM: general latent variabel modeling. Behaviormetrika 29: 81-117.
9. Muthen LK, Muthen BO (2004) Mplus User's Guide. (6th Edition) Los Angeles, CA. Muthén & Muthén.
10. Magidson J, Vermunt JK (2002) Latent class models for clustering: a comparison with K-means. Can J Marketing Res 20: 37-44.
11. Montana G, Pritchard JK (2004) Statistical tests for admixture mapping with case-control and cases-only data. Am J Hum Genet 75: 771-789.
12. Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 6: 21-45.
13. Oza NC, Tumer K (2008) Classifier ensembles: Select real-world applications. Information Fusion 9: 4-20.
14. Kuncheva LI (2004) Classifier ensembles for changing environments. In 5th Int Workshop on Multiple Classifier Systems, Calgliari, Italy.
15. Dorogovtsev SN, Mendes JFF, Samukhin AN (2002) Moderne architecture of random graphs: Constructions and correlations. arxiv:con-math 0206467.
16. Newman MEJ (2009) The structure and functions of complex networks. SIAM Review 45: 167-256.
17. Bianconi G, Pin P, Marsili M (2009) Assessing the relevance of node features for network structure. Proc Natl Acad Sci U S A 106: 11433-11438.
18. Koschützki D, Lehmann A, Peeters L, Richter S, Tenfelde-Podehl D, et al. (2005) Centrality indices. Network Analysis. Springer-Verlag Berlin Heidelberg.
19. Latora V, Marchiori M (2004) A measure of centrality based on the network efficiency. arxiv:con-math 0402050.
20. Lima-Mendez G, van HJ (2009) The powerful law of the power law and other myths in network biology. Mol BioSyst 5: 1482-1493.
21. Boguna M, Pastor-Satorras R (2003) Class of correlated random networks with hidden variables. Phys Rev E Stat Nonlin Soft Matter Phys 68: 036112.
22. Newman ME (2002) Assortative mixing in networks. Phys Rev Lett 89: 208701.
23. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69: 026113.
24. Karrer B, Newman ME (2011) Stochastic blockmodels and community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 83: 016107.
25. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E Stat Nonlin Soft Matter Phys 74: 016110.
26. Ronhovde P, Nussinov Z (2009) Multiresolution community detection for megascale networks by information-based replica correlations. Phys Rev E Stat Nonlin Soft Matter Phys 80: 016109.
27. Anand K, Bianconi G (2010) Gibbs entropy of network ensembles by cavity methods. Phys Rev E Stat Nonlin Soft Matter Phys 82: 011116.

Author Affiliation

Top

[1]Copenhagen University Hospital at Hvidovre, Department of Clinical Biochemistry, Genetics, and Molecular Biology, Kettegaard All 26, 2650 Hvidovre, Denmark