A toolbox for causally cohesive genotype-phenotype modeling

Jon Olav Vik[1], Arne B. Gjuvsland[1], Stig W. Omholt[2]

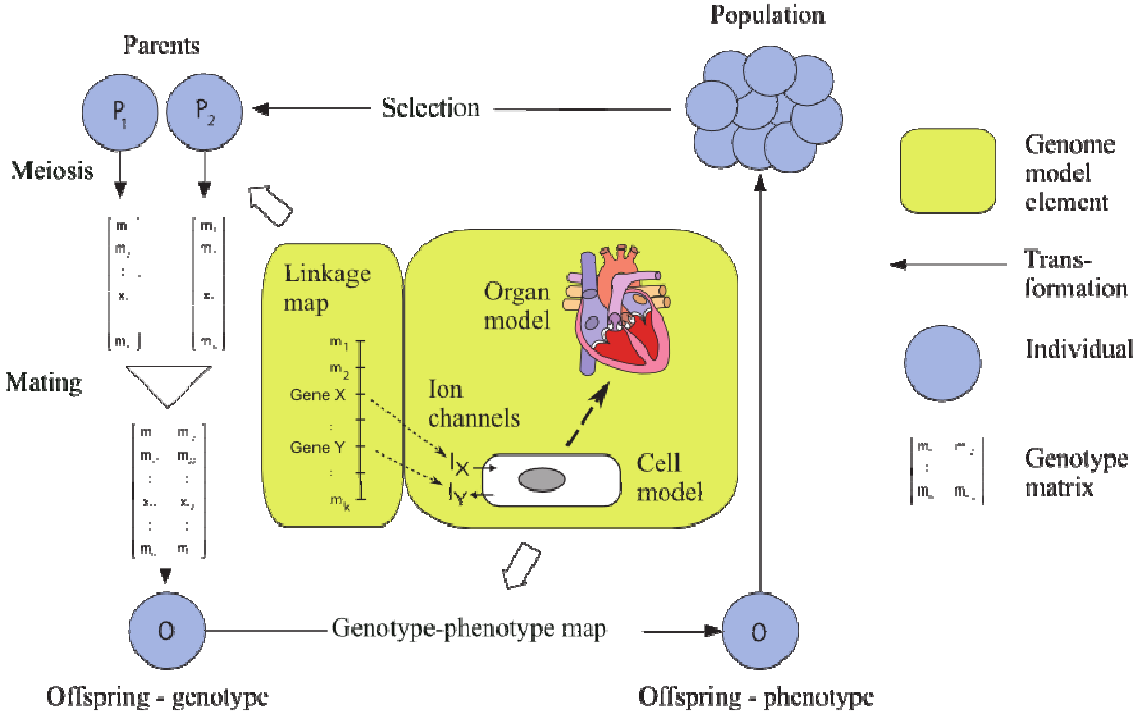
Norwegian University of Life Sciences, Ås, Norway

Correspondence: Jon Olav Vik, Centre for Integrative Genetics, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway. jonovik@gmail.com

A comprehensive understanding of how genetic variation causes phenotypic variation of a complex trait is a long-term disciplinary goal of genetics. Here we present a toolbox for utilizing curated mathematical models of physiology as a means towards this goal. The basic premise is that in a well-validated model that is capable of accounting for the phenotypic variation in a population, the causative genetic variation will manifest in the model parameters.

In this context, the term *phenotype* refers to any relevant measure of model behaviour, whereas the term *parameter* denotes a quantity that is constant over the time-scale of the particular model being studied. However, even the lowest-level model parameters are themselves phenotypes, whose genetic basis may be mono-, oligo- or polygenic, and whose physiological basis can be mechanistically modelled at ever deeper levels of detail.

We have proposed the term *causally cohesive genotype-phenotype modeling* (cGP modeling) to denote an approach where low-level parameters have an articulated relationship to the individual's genotype, and higher-level phenotypes emerge from the mathematical model describing the causal dynamic relationships between these lower-level processes (see figure below). It aims to bridge the gap between standard population genetic models that simply assign phenotypic values directly to genotypes, and mechanistic physiological models without an explicit genetic basis. This forces a causally coherent depiction of the genotype-to-phenotype (GP) map.



Integrating genetics with physiological models in a population setting. In the illustration, a gene codes for ion-channel parameters, which affect transmembrane currents and the action potential of a heart cell. Genetically determined variation in low-level parameters propagates through multiple levels of electrophysiological, mechanical and fluid dynamic processes. Phenotypic variation emerges at each level of organization. A cGP model integrates a multiscale model of this biological system with a linkage map through the genes encoding ion channels, thus the cGP model describes the creation of new genotypes as a result of meiosis and mating as well the phenotypes arising from these genotypes. By simulating populations of cGP models, whose parameters arise by recombination of virtual genomes, we can obtain a deeper understanding of the high-dimensional in silico phenotypic data.

The cGP approach has been further developed in an Exemplar Project of the Virtual Physiological Human Network of Excellence (VPH-NoE). The cgptoolbox (described below) has provided infrastructure for several applications, including the characterization of high-dimensional GP maps in relation to genetic concepts [1], high-dimensional sensitivity analysis in cGP models [2], and revitalizing genome-wide association studies by focusing on model parameters as informative phenotypes [3].

Outlook for personalized medicine

The Virtual Physiological Human Network of Excellence aims to apply multilevel physiological modelling in patient-specific healthcare and in simulation studies of disease-related processes. This requires expanding the scope of multilevel physiological modeling to the genome and population levels. Computational models of multi-level physiology imply a mapping from low-level parameters to clinically relevant phenotypes. Supplemented by a link from genomic databases to model parameters, this defines what we call a causally cohesive genotype-phenotype (cGP) model. Tailoring treatment to individual genetics is a stated goal of the VPH in the post-Genomic era[1]. However, understanding gene-disease associations requires population-level analyses accounting for genetic interactions and genotype frequencies in the population.

Aims of the cgptoolbox

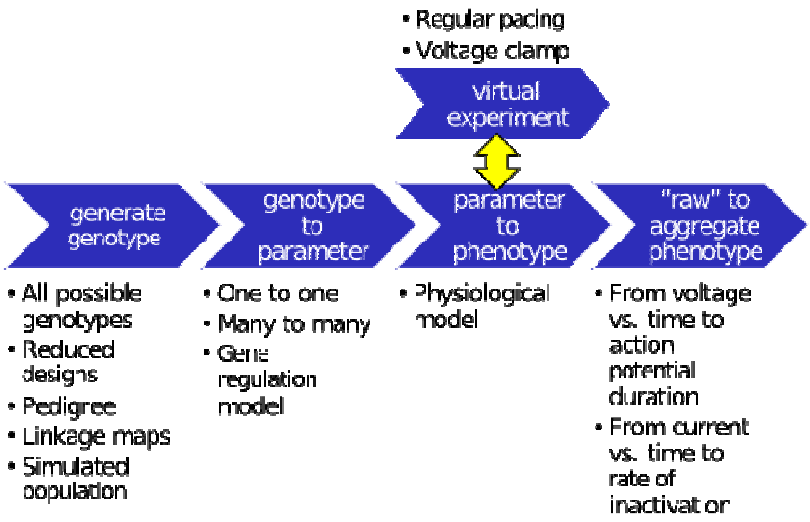
The cgptoolbox aims to facilitate researchers' entry into cGP modeling by providing a cGP modelling framework in a population context, integrating and interfacing with existing VPH tools. The toolbox will provide the means for extensive explorative *in silico* studies as well as integration of patient-specific information in multiscale models to account for the individual's genotype in the model parameterisation process.

The cgptoolbox is open source and hosted at https://github.com/jonovik/cgptoolbox. It adds to the VPH Toolkit by integrating genetic structure information, bioinformatic information and infrastructure, and multiscale and multiphysics models and associated infrastructure. The strength of the cGP toolbox as a relevant research tool will be illustrated by specific examples of use:

- as an explorative tool for better understanding of key genetic concepts like dominance, epistasis, pleiotropy, penetrance and expressivity in biologically realistic complex trait situations and in a patient-specific perspective;
- to elucidate the fine structure of the distribution of individuals in a high-dimensional phenotypic landscape associated with a pathological condition as a function of genetic variation;
- as a test bed for developing new fine mapping methodologies within statistical genetics aimed at exploiting high-dimensional phenotypic information.

The cgptoolbox is a step towards providing computational tools for attaching GP maps of parameters to a multiscale modelling framework in order to handle patient-specific issues. We think this is an important delivery preparing for a future situation where acquisition of high-dimensional phenotypic data from patients become routine [4] and the VPH community has come closer to its key goal of achieving more integration across multiple spatial and temporal scales.

Design philosophy



Simulation pipeline for causally cohesive genotype-phenotype studies. Blue arrows denote functions that generate genotypes or transform them through successive mappings, genotype to parameter to "raw" phenotypes to aggregated phenotypes. The surrounding text exemplifies different alternatives for each piece of the pipeline. *Virtual experiments* interact with physiological models to generate phenotypes defined by the system's response to external stimuli.

The workflow illustrated in the figure at left exemplifies the design pattern we developed to facilitate the interchange and reuse of its components: the generation of genotypes (e.g. exhaustive enumeration or reduced designs), the mapping of genes to parameters (based on genome databases, e.g. the mouse phenome project [5], physiological models (e.g. the CellML [6] and SBML/BioModels [7] repositories) that map parameters to phenotypes, virtual experiments to generate phenotypes that are defined by the model system's response to some stimulus or perturbation (e.g.voltage clamping), and aggregation from model dynamics to clinically relevant phenotypes (e.g. action potential duration). This pipeline design allows the gluing together of appropriate tools for each task. For instance, experimental designs and statistical analyses were done in R (www.r-

[1] http://www.vph-noe.eu/vph-repository/doc_download/13-vph-noe-promotional-flyer-v1

project.org), whereas virtual experiments were flexibly described in Python (www.python.org). The general approach should apply equally well to eventual whole-organ cGP studies.

cgptoolbox key components

- Using simuPOP [8] for **virtual genome data structure**, locating genes and markers on chromosomes, keeping track of physical and genetic map units, and providing slots for user-defined parameter data, which can be accessed by the physiological models. simuPOP also includes meiosis, taking into account chromosomal arrangement and recombination rates for both markers and functional genes, and functions for dealing with population structure and observed or model-generated pedigrees.
- Using Biopython to **import genomic data from public databases** (e.g. the HapMap project and Entrez databases such as SNP and Gene) into virtual genomes.
- **Core functionality for doing population-level simulations** combining structural genome dynamics (keeping track of recombination, allele frequencies and haplotype block structures) with cGP models (in addition to the traditional GP models from quantitative genetics). The software will be designed to be modular such that cGP models and pedigree structures can be easily changed. Examples will span the range from cellular models in CellML (see below) to whole-organ simulations of continuum dynamics using openCMISS.
- **Routines for turning CellML models into cGP models**. This will be done with as little manual work as possible, with automatic download from the CellML repository and integration using the CVODE solver.
- **Design patterns for virtual experiments** that interact with physiological models to generate phenotypes defined by the system's response to external stimuli. For instance, a given pacing protocol can be applied to a whole class of heart cell models.
- **Setting up simulations based on publicly available genomic data** from the HapMap project and Entrez databases such as SNP and Gene, using Biopython.
- **Export routines** to data formats for state-of-the-art quantitative genetic software for doing heritability estimates, haplotype block detection and genome-wide association studies.
- **Convenient packaging into tasks** that can be run trivially in parallel on computer clusters, automatically consolidating results as they become available.

References

1 Vik, J. O., Gjuvsland, A. B., Smith, N. P. & Hunter, P. J. 2011 Genotype–phenotype map characteristics of an in silico heart cell. *Frontiers in Physiology* **2**, 106. (doi:10.3389/fphys.2011.00106)

2 Tøndel, K., Indahl, U. G., Gjuvsland, A. B., Vik, J. O., Hunter, P., Omholt, S. W. & Martens, H. 2011 Hierarchical Cluster-Based Partial Least Squares Regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models. *BMC Systems Biology* **5**, 90. (doi:10.1186/1752-0509-5-90)

3 Wang, Y., Gjuvsland, A. B., Vik, J. O., Smith, N. P., Hunter, P. J. & Omholt, S. W. 2012 Parameters in dynamic models of complex traits are containers of missing heritability. *PLoS Comput Biol* **8**, e1002459. (doi:10.1371/journal.pcbi.1002459)

4 Houle, D., Govindaraju, D. R. & Omholt, S. 2010 Phenomics: the next challenge. *Nat Rev Genet* **11**, 855–866. (doi:10.1038/nrg2897)

5 Hancock, J. M., Mallon, A.-M., Beck, T., Gkoutos, G. V., Mungall, C. & Schofield, P. N. 2009 Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mamm Genome* **20**, 457–461. (doi:10.1007/s00335-009-9208-3)

6 Lloyd, C. M., Lawson, J. R., Hunter, P. J. & Nielsen, P. F. 2008 The CellML Model Repository. *Bioinformatics* **24**, 2122–2123. (doi:10.1093/bioinformatics/btn390)

7 Le Novere, N. et al. 2006 BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucl. Acids Res.* **34**, D689–691. (doi:10.1093/nar/gkj092)

8 Peng, B. & Kimmel, M. 2005 simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686 –3687. (doi:10.1093/bioinformatics/bti584)