Practical steps to digital organism models, from laboratory model species to 'Crops **in silico**'

Citation for published version:
Millar, AJ, Urquiza Garcia, J, Freeman, P, Hume, A, Plotkin, GD, Sorokina, O, Zardilis, A & Zielinski, T 2019, 'Practical steps to digital organism models, from laboratory model species to 'Crops *in silico*'', *Journal of Experimental Botany*, vol. 70, no. 9, JEXBOT/2018/235879, pp. 2403–2418. https://doi.org/10.1093/jxb/ery435

Digital Object Identifier (DOI):
10.1093/jxb/ery435

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Experimental Botany

Practical steps to digital organism models, from laboratory model species to ‘Crops *in silico*’.

Andrew J. Millar*[1], J. Uriel Urquiza[1], Peter L. Freeman[2], Alastair Hume[1,3], Gordon D. Plotkin[5], Oxana Sorokina[6], Argyris Zardilis[1], Tomasz Zielinski[1].

Running title: Realising digital plant models

[1]SynthSys and School of Biological Sciences, University of Edinburgh, EH9 3BF, UK.

[2] Roslin Institute, University of Edinburgh, Roslin, EH25 9RG, UK.

[3] EPCC, Bayes Centre, University of Edinburgh, 47 Potterrow, Edinburgh, EH8 9BT, UK

[5] Laboratory for the Foundations of Computer Science, School of Informatics, University of Edinburgh, EH8 9AB, UK.

[6] Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, UK.

* Corresponding Author. +44 131 651 3325

All authors’ ORCID, e-mail.

Uriel Urquiza 0000-0002-7975-5013, jurquiza@ed.ac.uk

Alastair Hume 0000-0001-8322-9182, a.hume@epcc.ed.ac.uk

Argyris Zardilis 0000-0003-4582-9421, a.zardilis@sms.ed.ac.uk

Tomasz Zielinski, Tomasz.Zielinski@ed.ac.uk

Gordon D. Plotkin 0000-0001-8496-6096, gdp@inf.ed.ac.uk

Oxana Sorokina, 0000-0002-6786-9945, Oksana.Sorokina@ed.ac.uk

Peter L. Freeman, 0000-0002-1496-9969, peterfreeman46@gmail.com

Andrew Millar 0000-0003-1756-3654, Andrew.millar@ed.ac.uk

Submitted   25 November 2018; Main text: 7700 words; 1 table; 5 figures.

Keywords: *Arabidopsis thaliana*; biochemical kinetics; community standards; computational modelling; data science; whole cell modelling.

Highlight [<30 words]

Combining models of biology across scales, for fundamental understanding and crop improvement, presents multiple challenges. We review practical experiences and promising approaches in the pursuit of digital organism models.

Abstract [198 words]

A recent initiative named "Crops *in silico*" proposes that multi-scale models "have the potential to fill in missing mechanistic details and generate new hypotheses to prioritize directed engineering efforts" in plant science, particularly directed to crop species. To that end, the group called for "a paradigm shift in plant modelling, from largely isolated efforts to a connected community" (Marshall-Colon *et al.*, 2017). 'Wet' (experimental) research has been especially productive in plant science, since the adoption of *Arabidopsis thaliana* as a laboratory model species allowed the emergence of an Arabidopsis research community. Parts of this community invested in 'dry' (theoretical) research, under the rubric of Systems Biology. Our past research combined concepts from systems biology and crop modelling (Chew *et al.*, 2017; Chew *et al.*, 2014b). Here we outline the approaches that seem most relevant to connected, 'digital organism' initiatives. We illustrate the scale of experimental research required, by collecting the kinetic parameter values that are required for a quantitative, dynamic model of a gene regulatory network. By comparison to the SBML community, we note computational resources and community structures that will help to realise the potential for plant systems biology to connect with a broader crop science community.

Introduction

What distinguishes crop modellers from systems biologists, one of us was told ten years ago, is some responsibility to feed the world population. Systems Biology aims to understand the interactions among the component parts of a living system and the emergent properties that arise from such interactions (Alberghina and Westerhoff, 2005; Kitano, 2002). Its aspiration was to include components across multiple scales from the molecular to at least the organism. In practice the research started from intracellular pathways and only gradually intersected with physiological, organism-level approaches; most often, the organism in mind was a human (Kitano, 2015). Readers seeking to pin down systems biology, to a claim for novelty or otherwise, should consult earlier commentaries (Bothwell, 2006; Hammer *et al.*, 2004; Marcum, 2008). The holistic, systems approach led to a meeting with mission-orientated research in crop science, though the whole-plant scale to which Systems Biology aspired was then at the lower bound for crop models. The approach also distinguished Systems Biology from much research focusing on the properties of individual, biological components.

Along with the move from reductionism towards holism came a need for the ‘dry’ methods of formal modelling, because the unaided human brain is quite inept in reasoning quantitatively about dynamical systems as complex as those in biology. Several areas of plant science (cell physiology and ecology, to name but two) and crop science, have been ‘amphibious’ for decades, mixing ‘wet’ (experimental) and ‘dry’ (theoretical) approaches. The benefits of interfacing plant systems biology with crop modelling were recognised over a decade ago (GARNet Advisory Committee, 2006; Thomas, 2007), not only for modelling expertise but also for the real-world impacts. Crop models are regularly used by growers, breeders and Earth scientists, amongst others. Ten years later, an initiative named “Crops *in silico*” proposed that multi-scale models “have the potential to fill in missing mechanistic details and generate new hypotheses to prioritize directed engineering efforts” in plant science, particularly directed to crop species. To that end, the group (including A.J.M.) called for “a paradigm shift in plant modelling, from largely isolated efforts to a connected community” (Marshall-Colon *et al.*, 2017; Zhu *et al.*, 2016). However, formal models have been largely absent from the training of plant biologists, so this seemingly-natural interface has emerged only slowly. The diversity of models may also be less obvious for plant researchers, though it is arguably as great as the diversity of experimental methods. Crops *in silico* aims to link several, current approaches, such as functional-structural plant models that have organ-scale spatial resolution and process-based crop models with lower spatial resolution.

Dealing with diverse models is inevitable in the holistic agenda of Systems Biology. This article outlines some types of model that seem valuable for a community initiative such as “Crops *in silico*”. Our experiences, tools and approaches to combine and use them arose particularly from joint work on the Framework Model for Arabidopsis growth (Chew *et al.*, 2017; Chew *et al.*, 2014b), which in part followed practices from crop modelling. *Arabidopsis thaliana* emerged as the laboratory model species for plant science, with an open research community (Ankeny and Leonelli, 2011; Leonelli, 2007), about fifteen years before Systems Biology emerged as a research field (Vermeulen, 2017). We illustrate results, resources and social organisation of Arabidopsis research that are benefitting plant Systems Biology, and could further contribute to and benefit from the interaction with crop science. The challenge is to ensure that actual researchers with particular skill sets are motivated and able to complete research in realistic time, and to make the results comprehensible, useful and reproducible for others. We point to current, computational tools and resources that will help to realise this potential.

Standpoint

The authors represent a spectrum of systems biology research, spanning plant science, molecular biology, computer science, research management, software engineering and advanced computation. We are linked by research in or associated with SynthSys, the centre for Synthetic and Systems Biology at the University of Edinburgh, which has a long association with Systems Biology (Bard, 2008) and with Science, Technology and Innovation Studies in social science (Henry, 2008). A.J.M. previously coordinated GARNet, the UK community organisation for Arabidopsis researchers (see Box 1) and contributed to the "Crops *in silico*" proposals.

The diversity of "models"

A biologist's "model" often describes the contemporary understanding of a biological process, expressed in text, or as a diagram or cartoon (Figure 1A). Such descriptions are informal and very useful as a distillation of biological knowledge, but they are fatally flexible, ultimately ambiguous and difficult to reuse in a formal context. In contrast, mathematical models are formal and unambiguous, inflexibly imposing a rigour of description that often exposes serious gaps in biological knowledge. Identifying such gaps can be extremely valuable to direct ongoing work but the gaps must be bridged with assumptions in order to complete a model.

We summarise below some modelling approaches used in Systems Biology, based broadly upon their explanatory ability. An explanatory model can illuminate the mechanisms of a biological system and its principles of operation, whereas a descriptive model simply aims to predict the behaviour of the system based upon its past behaviour, irrespective of the biological mechanisms. Models in crop science and in systems biology each span this range. Models of "Crops *in silico*" will usually combine several approaches, so more detailed classification is difficult (Coveney and Fowler, 2005). Rather, we highlight opportunities for each model type in building complex models in plant and crop science. Detailed spatial models of plant development have been reviewed elsewhere (Ndour *et al.*, 2017; Prusinkiewicz and Runions, 2012; Truskina and Vernoux, 2018). Despite omitting this area for brevity, we note that models of cellular processes at the shoot apical meristem (Jonsson *et al.*, 2006; Kierzkowski *et al.*, 2012) or in lateral root formation (Dyson *et al.*, 2014; Xuan *et al.*, 2016) have often combined multiple model types.

Graphical models

A useful, formal description of a biological process can start without equations or computer programming, because a diagram can be formal (as can a text description). A defined vocabulary of graphical symbols (glyphs) can represent the various types of biological components as nodes in the diagram, with a defined set of connecting arcs to represent the processes by which the components interact. Drawing such a diagram can reveal gaps in understanding and record the assumptions made to bridge the gaps, as noted above. Maps of the metabolic network are a familiar example but complex models need to represent much more than metabolism. The Systems Biology Graphical Notation (SBGN) is a community standard for drawing intracellular pathways (Le Novere *et al.*, 2009), representing various types of molecules, their modifications, complexes, compartments and so on. SBGN is supported by free software tools, such as VANTED (Rohn *et al.*, 2012) and Cytoscape (Goncalves *et al.*, 2013). These can be extended to support other notations, for example for plant structures. Several online repositories provide SBGN diagrams of pathway information or models for download (Buchel *et al.*, 2013; Naithani *et al.*, 2017). A diagram of this type can comprehensively represent the state of knowledge, as a valuable addition to a review publication. A hand-curated diagram of mTOR response pathways included 964 molecular components, for example (Caron *et al.*, 2010) but such a large diagram is difficult to read in practice. Moving from a diagram to a quantitative model requires additional stoichiometry and parameter values, which can be added in graphical modelling software such as Cell Designer (Funahashi *et al.*, 2008) and Simile (Muetzelfeldt and Massheder, 2003).

For a diverse and growing community like “Crops *in silico*”, investing in graphical models offers three advantages. A non-modeller should be able to find, download and start to modify an existing diagram to represent their process of interest within 30 minutes, without prior preparation. This is the fastest route to modifying a model, similar in approach to the graphical languages used to teach computer programming (Marji, 2014). An expert modeller could use such a diagram as a starting point for detailed modelling of an unfamiliar process, similar to the pseudo-code that is used to sketch software functions prior to full coding. For experts and non-experts alike, the diagrams also offer a human-readable format to orient themselves quickly within a model.

Data-driven modelling

High-throughput technologies such as automated phenotyping platforms capture information on many components of a system simultaneously. Analysis of high-throughput data involves modelling with statistical techniques such as clustering, principal component analysis (PCA) and regression (Jagaman and Danuser, 2006). Similar methods can apply to the meta-analysis of data curated from the literature (Poorter *et al.*, 2012), with very broad scope (Diaz *et al.*, 2016). These data-driven methods can use little or no prior knowledge about the system and overlap with the expanding range of machine learning approaches, such as neural networks (reviewed in Ma *et al.*, 2014). Data-driven methods are usually descriptive and can inform simple, mathematical relationships that are used in many models where more detail is unavailable or undesirable. They represent a relevant process concisely, in sufficient detail to lead to the formation of specific hypotheses, for example about the mechanisms that underlie the differences between clusters (Janes and Yaffe, 2006) or the connections among variables (Dalchau *et al.*, 2011; Onoda *et al.*, 2017). Thus advanced analysis by data-driven methods grades into conceptual modelling (Valladares *et al.*, 2014). In a spatial context, Mundermann et al. (Mundermann *et al.*, 2005) modelled the development of the Arabidopsis shoot in the L-studio software, using measurements of architectural parameters to support detailed simulation and realistic visualisation of plant growth (Figure 3).

The articles by Dalchau et al. and Mundermann et al. used data generated by the same labs that conducted the modelling, which is common in small or emerging fields that use laborious assays. In contrast, the work of Poorter and colleagues allows meta-analysis of many data sets from well-established, eco-physiological assays (Poorter *et al.*, 2010). The more data is required for a modelling project, the more data availability can limit its progress and the career prospects of the modellers. The Open Research movement, with its FAIR and Open data principles, deserves their wholehearted support (see final section).

Baker *et al.* (2018) argue that data-driven methods' rapid focus on results may be more attractive for research that is close to professional practice (clinical medicine in their case), whereas other disciplines emphasise explanatory power. Several benefits can clearly follow from integrating these approaches. Our work on the circadian clock encountered some practical difficulties in this process. Data-driven approaches to learn the gene circuit structure were hampered by the very non-linearity, time-dependency and density of interactions that had originally motivated us to initiate modelling studies, remaining difficult even with a series of new methods (for example, Aderhold A., 2013; Grzegorczyk *et al.*, 2008; Higham and Husmeier, 2013). In contrast, data-driven connections of the clock to metabolism were published (Grzegorczyk *et al.*, 2015) and personnel had moved on, years before the follow-up experimental studies were complete (Flis *et al.*, 2015; Flis *et al.*, 2018).

Qualitative modelling

Whereas data-driven models can represent detailed data with little explanatory power, qualitative models offer explanatory power with limited detail. Boolean models are the most common type, where components and connections are represented as present or absent, and this coarse state of the system may change over time. These models test hypotheses about the logical and causal relationship between events, stimuli and system responses (De Jong, 2002). An early example in plant science represented the network of transcription factors that specify organ identity during Arabidopsis flower development. The model's logical rules tested (and supported) the conceptual "ABC model" of gene interactions (Espinosa-Soto *et al.*, 2004). Complex waveforms can be represented by allowing a time delay between the activation of one component and the next, yet the models remain attractively concise. A time-delay model (Figure 2) allowed us to test all possible connections among the genes of the Arabidopsis circadian clock (Akman *et al.*, 2012), for example, highlighting a new circuit that explained the experimental data better than the circuit proposed at the time. This qualitative model's circuit was independently confirmed by new data and in a more detailed, quantitative model from our lab (Pokhilko *et al.*, 2013). Note that we could not have tested all possible circuits in the quantitative model in a reasonable computation time.

For Crops *in silico*, Boolean models (and other qualitative models) might be the easiest way to incorporate large gene-regulatory networks. They do, however, risk discarding information for the best-studied components, which may have sufficient data for more detailed treatment. Hybrid models are then natural, where some components are represented in qualitative and others in quantitative form. For example, a binary representation of (unmeasured) transcriptional activation of a reporter gene allowed us to test several possible gene circuits in an algal clock, combined with a continuous, quantitative model for the levels of a luminescent reporter protein that reproduced experimental data (Ocone *et al.*, 2013). The software to support logic models is growing, exemplified by development of the Systems Biology Markup Language (SBML) "qual" standard for model exchange (see below)(Buchel *et al.*, 2013). Software tools can also help in converting qualitative models to quantitative forms (Wittmann *et al.*, 2009), which is not yet a common path (Ortiz-Gutierrez *et al.*, 2015) but might become a natural progression for Crops *in silico* as more data becomes available (Le Novere, 2015).

Constraint-based modelling

Even dynamic, biological systems can be treated as being in steady state, when their homeostatic mechanisms buffer changes, at least substantially. The numbers of some molecule being generated and degraded are equal, for example, so its level is almost constant in time. Additionally, the time scale for metabolic events (seconds) is typically much faster than for genetic regulation (hours): from the perspective of genetic regulation, the metabolic system is always in steady state. The characteristics of this constant state depend on the structure of the system (the related biochemical reactions and their stoichiometry), general thermodynamics laws and external parameters, such as the cellular energy supply. Where a metabolic network is well understood, for example, constraint-based analysis is able to identify a set of fluxes through the network that are compatible with the observed steady state, to predict missing reactions and alternative pathways, and to find steady states that become accessible under different conditions. More prior knowledge is required than for qualitative models, and the models have greater explanatory power. In the areas relevant to Crops *in silico*, De Reuille et al. used constraint-based modelling to create the geometry of the shoot apical meristem, subsequently using this geometry as a constraint for auxin transport to evaluate the distribution of auxin fluxes (Reuille *et al.*, 2006). The approach can be extended to represent data that change over time, such as day and night states of central carbon metabolism (Cheung *et al.*, 2014) or the hourly dynamics of the starch pathway (Sorokina *et al.*, 2011). These extensions for dynamic systems are limited and development is ongoing. They are attractive in principle for Crops *in silico*, because constraint-based models are computationally tractable and do not require the detailed kinetic parameters of full, quantitative models.

Quantitative modelling

Quantitative modelling techniques represent the most detailed explanation of the underlying mechanisms and allow the most extensive numerical comparison of simulation results with experimental data. Correspondingly, they require the most prior information on the system (illustrated below). Where changes over time (dynamics) are of interest in the biology, for example in the cell cycle or the circadian clock, these methods have given impressive results (Bujdoso and Davis, 2013; Novak and Tyson, 2008; Tyson and Novak, 2015). Systems of ordinary differential equations (ODE) are a popular approach where time is continuous, as are the equivalent, difference equations with discrete time steps. Each equation describes the change in one variable (organ mass, protein concentration etc.) as a sum of reactions (synthesis, destruction, transport etc.) that are represented with empirical, kinetic terms (law of mass action, Michaelis-Menten approximation, piecewise-linear functions etc.). Variables can justifiably be continuous, implying an infinite number of intermediate concentrations, if molecular numbers are in fact large, reactions are frequent and the system behaves reproducibly. This style of modelling is common in plant Systems Biology and has been reviewed elsewhere (Chew *et al.*, 2014a; Middleton *et al.*, 2012). However, data at the single-cell level increasingly reveals components that are present in small numbers (Libault *et al.*, 2017), where the continuous, deterministic approach is inaccurate and instead discrete, stochastic models describe the probabilities of each reaction event (Shahrezaei and Swain, 2008). Stochastic models of the plant clock circuit suggested that circadian timing would be variable at the single-cell level, for example (Guerriero *et al.*, 2012), as recently confirmed experimentally (Gould *et al.*, 2018). Multi-model frameworks like Crops *in silico* must therefore anticipate stochasticity at this micro-scale, in addition to the formation of discrete organs in a plant model, or germination of individual weeds in a field model.

Multiple types of model are as natural in a digital organism as the many biological processes that contribute to a physical organism (or the many research perspectives to understand it). Integrating these diverse model types is by no means only a technical topic. In the example of data-driven and quantitative modelling approaches to the circadian clock (above), flexible management was required (Balmer *et al.*, 2016) to reconcile the timelines of each modelling approach and their different concepts of the "publishable unit" of research. New approaches to research dissemination could be adopted in a Crops *in silico* community, as preprints, data publications, model archive files, and institutional innovations such as "inside-out" libraries (Bergmann *et al.*, 2014; Dempsey, 2013; Leitner *et al.*, 2016; Schloss, 2017) offer more flexibility in what constitutes a "unit" for dissemination. We return to these social factors in the context of community standards, below, and in the final section.

Modelling frameworks and languages

The technical challenge to link heterogeneous models is long-standing and well recognised (Adam *et al.*, 2012; Ghosh *et al.*, 2011; Goldberg *et al.*, 2018; Macklin *et al.*, 2014; Marshall-Colon *et al.*, 2017; Pradal *et al.*, 2008). The approaches can be simplified to two extremes, either to rewrite all the models in a common modelling language or to devise an integration system that links the models in their diverse, native forms, as loosely-coupled "black boxes" (Figure 3). Tightly woven into this problem is the distinction between declarative and procedural models. Declarative models are a formal specification of the model, such as its mathematical definition. Separate software is then required to simulate the model, leading to advantages described elsewhere (Muetzelfeldt, 2007). If in addition a declarative model uses a standardised format, then the model becomes easy to exchange between software tools (discussed in the following section), and therefore easier to understand and modify.

In contrast, implementing the model in a programming language is procedural (or ‘imperative’): the model specification is also the computer code for simulation, whether it is in a scripting language such as python or R, a high-level language such as Matlab, or a general-purpose language such as C++. Good programming conventions can separate the declarative part of the model but there is no guarantee of this. The code may then be executable but obscure, making the model a black box. Modelling procedures are clearly important as well as the models. Open-source, well-documented code makes these more accessible than a closed-source or undocumented modelling framework. The importance of open-source software for reproducible research is discussed elsewhere (Mendes, 2018).

To illustrate these general considerations with a detailed example, we consider the development of the Arabidopsis Framework Model from four previously-separate models (Chew *et al.*, 2014b). Rewriting each of the constituent sub-models into a common language in the Simile modelling environment, then re-validating them in numerical simulation, was a major effort (Muetzelfeldt and Massheder, 2003). A preliminary project, PlaSMo, first collected likely component models from idiosyncratic computing code (Davey *et al.*, 2009). The refactoring process depended on access to the model files. Files for one model had been deleted online and were only available from the Google cache. The commercial, Simile environment was selected for refactoring because it offered a rich, graphical interface and supported a declarative, XML model format, SimileXMLv3 (see Box 1). Like SBML, this was based on the widely-used MathML standard (Hucka *et al.*, 2003). In practice, refactoring the various model codes required unusually broad skills. As benefits of this investment, the component models in a web portal (see Box 1) became more readily and uniformly accessible for future work, and the process of model curation and re-validation provided stringent quality control. Among the challenges were IF … ELSE … conditions: standard programming tools, which might distinguish parts of a model that are used at different stages of plant development. These effectively, and very concisely, embed multiple, alternative models within the same procedural code. Rewriting such models could involve untangling a web of conditional statements, improving clarity but expanding the model description. The Agricultural Model Exchange Initiative (Martre *et al.*, 2018) are currently embarking on a similar approach, with contemporary software tools (see Box 1).

The "black box" approach is initially faster, at least for a small number of models. The L-studio framework, for example, can call external model codes (Figure 3), and the emerging Crops *in silico* interface links models in four programming languages (see Box 1). More ambitious model integration systems have been applied in projects (Marshall-Colon *et al.*, 2017; Zhu *et al.*, 2016) such as the European agricultural assessment project SEAMLESS (van Ittersum *et al.*, 2008). The promise of this loose coupling is that modellers continue to develop their diverse, component models independently, and yet can still interact with the ensemble. The practical risk is that their unencumbered innovation flies beyond the reach of the integration system, so the ensemble can no longer be simulated. More dangerously for the long term, a growing set of 'black box' models is harder for any individual to understand, frustrating the need for modellers to refine and revise the component models. This seems to be an opportunity for biology to inspire new computer science, for example using domain-specific languages that naturally express the relevant biology (Honorato-Zimmer *et al.*, 2017; Kniemeyer *et al.*, 2007; Zardilis *et al.*, 2019) and meta-languages that integrate these models and control their simulation (Mjolsness, 2018).

Standards-based modelling for Crops in silico

If a growing number of plant modellers are to understand and use a wider range of model types, investing in a standards-based approach can speed up the process. Systems Biology uses several modelling standards, notably Systems Biology Markup Language (7) and Cell Markup Language (CellML). SBML is a standard for constraint-based and quantitative models (Hucka *et al.*, 2018). CellML adds support for various cellular interactions (Lloyd *et al.*, 2004). These machine-readable, model exchange formats (Figure 1C) that have spurred investment in a mutually-reinforcing economy of online repositories and software tools that use the standard format as input and/or output. For example, storing a private SBML model file in the self-service FAIRDOM data repository (Wolstencroft *et al.*, 2015) automatically allows simulation of the model at the JWS-online resource (Snoep and Olivier, 2002). Complementary standards are growing the economy. The Simulation Experiment Description Markup Language (SED-ML), for example, describes how a particular SBML model simulation was run (Waltemath *et al.*, 2011). Uploading a SED-ML file to an online resource can exactly reproduce a published simulation figure. The file specifies how the resource should retrieve a model file from an online repository, send it to an online simulator and plot the relevant part of the simulation results. This level of transparency and replicability is a highly attractive product of the global SBML economy (Mendes, 2018). Given these potential advantages, we considered how SBML would represent a plant growth model that might arise from Crops *in silico*.

The plant growth use case highlighted three main issues for SBML: input weather data, expressing some key concepts, and simulators for multi-models. First, systems biology models usually reflect controlled, laboratory conditions. The Input Signal Step Function in SBML represents step and cyclic experimental manipulations (Adams *et al.*, 2012), for example, motivated by the light-dark cycle in a plant growth chamber. Most crop models, in contrast, read in timeseries of fluctuating weather data during the simulation. SBML does support custom-defined functions, including splines and piecewise-linear functions. These can represent input timeseries data as new variables in the SBML model file, interpolating between timepoints to make environmental data available at any point in the simulation. Simple SBML Data Tools were therefore created to support such modification of SBML files, for crop and other models (see Box 1). Secondly, core SBML cannot represent the creation of compartments during a simulation, as required to model the formation of new plant organs. SBML development was revised in 2010 to extend the core (Hucka *et al.*, 2018) with specialised, modular packages, which are proposed by the community ("qual" was noted above). Three packages were particularly relevant for the Arabidopsis Framework Model, which would be representative for many plant-level models: arrays, dynamic processes (the package known as "dyn") and hierarchical model composition ("comp"), among a larger set that was discussed earlier (Muetzelfeldt, 2010). Productive interaction with any such community effort needs some understanding of the community norms. The packages are at varying stages of development (SMBL community, 2017). SBML community rules focus their resources on the exchange of models between software tools, where there is demand for the exchange and support for its standardisation (Hucka *et al.*, 2015; Schreiber *et al.*, 2015). To be formally adopted, new SBML packages must be implemented in two, independent software products. A potential drawback of the modular approach is that, even if each of the three packages mentioned is fully developed in SBML, there is no guarantee that any simulation software will support all three together. Engaging with SBML models offers a bridge to Systems Biology but the sensible norm that demand and software tools together lead the development of SBML standards, as noted above, has a significant repercussion. Both demand and tools will initially be limited, when an initiative such as Crops *in silico* aims to lead a field. Engagement with community standards might therefore be a later step. Lastly, controlling disparate simulation timesteps and reconciling the availability of shared resources among competing sub-models were considered at a workshop in 2015, which tested the representation of a landmark "whole cell" model (Karr *et al.*, 2012) in a standardised form (Waltemath *et al.*, 2016). One option considered for modular, multipart models was a model-control system, using a standard akin to SED-ML. This approach might be equally relevant to integrating diverse models for Crops *in silico*. However, the workshop report coyly notes that "Significant effort will also be needed to develop an efficient, parallelized, multi-algorithm simulator." (Waltemath *et al.*, 2016).

After a suitable modelling approach has been selected, the modellers must represent the biological processes of interest with enough detail to address the relevant issues. The question of "what's in the model" (specifying the model's variables) usually has many reasonable answers, which provoke debate rather than consternation. If the biological issues require a quantitative model, however, specifying the rates that are associated with each process (the values of the model's parameters) can be an overwhelming and contentious task. We next provide a specific example that illustrates this challenge.

Parameter values for a quantitative model

The 24-hour, circadian clock in *Arabidopsis thaliana* has been a paradigmatic system for studies of dynamic gene regulation over 20 years (Millar, 2016). Because timing was the critical, biological issue, quantitative, dynamic models were a natural approach (Bujdoso and Davis, 2013). They operated with time in real hours and their success was judged on whether the simulated waveforms of rhythmic gene expression helped to understand (explain and predict) the experimental timeseries data, in various conditions. The RNAs and proteins of the dozen or so clock genes were represented with arbitrary concentration units, in contrast to the real hours. These models were built to understand results from molecular genetic assays, which often uses relative or arbitrary units, rather than biochemical kinetics, where absolute units are more common. Models in absolute units are advantageous, however (as outlined below). We therefore summarise the parameter values that would be required to convert a model of a plant gene regulatory network, such as P2011 (Pokhilko *et al.*, 2012), to absolute concentration units. The values described are listed in Table 1, extending similar resources of parameter estimates for other organisms (Milo *et al.*, 2010).

Macromolecular synthesis and degradation

Most of the models deal with the birth and death of the clock gene RNAs and proteins. However, absolute RNA transcription rates have not been measured in plants. Sidaway-Lee et al. (Sidaway-Lee *et al.*, 2014) measured the distribution of nucleotide incorporation rates in

Arabidopsis and their temperature-dependence. The results were reported in microarray fluorescence units per hour. We are therefore limited to estimating a maximum transcription rate for eukaryotes in general, from a maximum RNA polymerase II elongation rate of 5 kbp/minute in human cell lines (Danko *et al.*, 2013) and 4.5 kpb/min in zebrafish (Hanisch *et al.*, 2013), and occupancy of typically one RNA polymerase complex per gene (Zenklusen *et al.*, 2008). Maximal transcription rate is then 2min^{-1} for a 2.5kb RNA, for example, ignoring short-term transcriptional bursting (Harper *et al.*, 2011). RNA degradation rates have been measured in large-scale studies (Narsai *et al.*, 2007; Sidaway-Lee *et al.*, 2014), either after transcriptional inhibition or by inference from the nucleotide incorporation data. Mean RNA half-life was 5.9h in plant cell cultures at 22°C (Narsai *et al.*, 2007), or 1.9h (at 27°C) to 5.0h in plants (17°C, Sidaway-Lee *et al.*, 2014). The microarray readout signals were less reliable for rare and unstable RNAs, however, and RNAs with daily rhythms must be unstable. Specific analyses of clock-relevant RNAs are therefore important, again using inhibitors (Lidder *et al.*, 2005) or by inference from statistical timeseries models without inhibition (Finkenstadt *et al.*, 2008). Note that the inhibitors could give paradoxical results (Finkenstadt *et al.*, 2008): if the degradation of a target RNA is regulated by an RNA mediator that is itself unstable, then rapid depletion of the mediator during a transcriptional block may stabilize the target RNA.

Protein translation rates were measured by Piques et al. (Piques *et al.*, 2009) for a set of metabolic-related genes in Arabidopsis, using calibrated qRT-PCR assays to measure the absolute number of transcripts in free RNA or bound to ribosomes. The fraction of transcripts engaged in translation can be calculated, yielding a range of 0.56-0.9, mean 0.77. A ribosome translation velocity of 3 amino acids/second and density of 6.6 ribosomes/kb of coding sequence (CDS), based on data from bacteria (Brandt *et al.*, 2009) were then used to estimate protein synthesis rates (mol protein g^{-1}FWh^{-1}) and their increase in the light compared to the dark period (Ishihara *et al.*, 2015; Piques *et al.*, 2009). Protein degradation rates have been measured in large studies following metabolic labelling (Li *et al.*, 2017), though the mass spectrometry methods involved are biased towards abundant and therefore often stable proteins and the dynamics of amino acid pools introduce further limitations (Ishihara *et al.*, 2015). The median half-life of 6 days (Li *et al.*, 2017) clearly does not represent the clock regulators with high-amplitude, daily rhythms. However, constraints on the possible protein degradation rates can be estimated from the available timeseries data, where the clock protein has been detected as a tagged fusion protein or with antibodies to the native protein (for example, Knowles *et al.*, 2008; Nakamichi *et al.*, 2010).

Volume and transport

Given these synthesis and degradation rates, various models can estimate molecular copy number per cell. The next critical values are the volumes of the relevant cellular compartments, to convert copy number estimates to concentrations. Koffler et al. (Koffler *et al.*, 2013) quantified the volumes of *A. thaliana* mesophyll cells in young and old leaves, reporting each compartment as a fraction of total cellular volume. For example, the mean volume occupied by the nucleus was 0.16% of the cell volume in an older leaf. Wuyts et al. (Wuyts *et al.*, 2010) report the distribution of volumes for palisade mesophyll cells, with a mean cell volume of 73,000μm^3. Combining these gives a nuclear volume of 117μm^3. This is reassuringly close to an estimate of 113μm^3 that we calculate from the nuclear diameter of 5.99 ± 0.72μm measured by 3D-FISH (Tirichine *et al.*, 2009), assuming a spherical nucleus.

Finally, model components must be transported among cellular compartments; in our case the nucleus is particularly relevant. No data is present for the size, number or distribution of *A. thaliana* nuclear pore complexes (NPCs), the route for such transport. Data on tobacco BY-2 cell cultures showed around 50 NPCs per μm^2 of nuclear envelope (Fiserova *et al.*, 2009). Furthermore, in human cultured HeLa cells the transport rates of NTF2 and Transportin are 170 and 140 molecules/s/NPC respectively (Kubitscheck *et al.*, 2005). If we assume that similar transport rates are achievable in *A. thaliana*, using the nuclear diameter above suggests possible transport rates up to 960,000 molecules/s into the nucleus. These are unlikely to affect dynamics on a circadian timescale of multiple hours, unless nuclear transport is specifically regulated.

Binding affinity

Clock proteins function in the model by interacting either with each other or with the DNA in a clock gene's promoter. The affinity (K_d) of each interaction affects the model's behavior but almost none of the specific values have been measured. General (Kastritis *et al.*, 2011; Kumar and Gromiha, 2006) or more specific (Stiffler *et al.*, 2007) databases describe protein-protein interactions in other species. Wide variation in even the median K_d (233nM, 12nM and 14μM, respectively) in part reflects the inclusion of protein classes such as high-affinity antibodies, emphasizing the importance of more targeted resources. A sample of 42 published DNA-protein affinities for plant DNA-binding proteins gives median K_d of 20nM (Figure 4A) (Aggarwal *et al.*, 2010; Hao *et al.*, 1998; Hofr *et al.*, 2009; Izawa *et al.*, 1993; Liang *et al.*, 2008; Moyroud *et al.*, 2009; O'Neill *et al.*, 2011; Prouse and Campbell, 2013; Reymond *et al.*, 2012). A similar collection of plant protein-protein interactions (n=45) suggested a median K_d of 86nM (Figure 4B) (Ballut *et al.*, 2005; Bauer *et al.*, 2013; Bernal-Bayard *et al.*, 2014; Bisson and Groth, 2010;

Dong *et al.*, 2010; Fuglsang *et al.*, 2003; Hao *et al.*, 2011; Levskaya *et al.*, 2009; Li *et al.*, 1999; Liu *et al.*, 2007; Luoni *et al.*, 2006; Mantovani *et al.*, 2014; Ogawa *et al.*, 2008).

Means and ends of detailed models with absolute parameterisation

One advantage of a model species such as Arabidopsis is the concentration of research effort, resulting in measured values for parameter such as the nuclear volume (above). Nonetheless, building a quantitative model of a plant gene regulatory network such as the P2011 clock model seems to demand more parameter values than have been measured. Parameter fitting is one means to overcome the incomplete parameter measurement, and was used extensively to construct past clock models (Bujdoso and Davis, 2013). Rather than being constrained by input parameters alone, the model outputs were constrained to match functional data, in this case the detailed waveforms of rhythmic timeseries. The data in Fig. 1D would help to constrain the clock model, for example. Timeseries data have been published by many research groups for tens of light-dark conditions and clock-mutant plants. Each timeseries typically has 10-100 data points. Public, reference data sets are available (Flis *et al.*, 2015), only for Arabidopsis, to ease the burden of data collation (Fogelmark and Troein, 2014). Mathematical analysis suggests that the clock might be particularly tractable to parameter fitting, because the interlocked, negative-feedback loops of gene regulation constrain the system’s dynamic behaviour (Rand *et al.*, 2006). Regulatory networks of this form have much less flexible behaviour than a modeler might expect to gain from the many parameters, so correspondingly fewer sets of parameter values can produce model outputs that match the timeseries data. Indeed, detailed measurements in Arabidopsis have subsequently validated some of the fitted parameter estimates of clock models (Pudasaini *et al.*, 2017), suggesting that more such measurements could further validate the approach.

Model development still required searching a high-dimensional space (several 10’s of parameters) to discover sets of parameter values that were consistent with the data, which is computationally demanding. We have shown that open data, free software (Alves *et al.*, 2006) and public computational resources can make this process accessible (Flis *et al.*, 2015) but experts in advanced computation will remain important contributors to Crops *in silico*. Absolute parameter estimates (above) are valuable here too, in limiting the range of values that the search algorithms must explore, speeding the parameter search. Moreover, qRT-PCR assays calibrated to absolute RNA copy numbers are now providing the first gene expression timeseries data that naturally match the simulation outputs from models with absolute parameter values (Baudry *et al.*, 2010; Flis *et al.*, 2015; Piques *et al.*, 2009).

Modelling with absolute biochemical units should benefit our understanding of the clock, judging by earlier examples in biology. We should discover whether the models' arbitrary units concealed some processes that required unusual or impossible parameter values, suggesting that the plant uses a different biochemical mechanism to achieve that aspect of its circadian timing. Unrelated studies (including high-throughput surveys) will more easily test parts of the model, by measuring a relevant biochemical parameter value or the level of a model component, compared to the model's predicted value (as noted above, Pudasaini *et al.*, 2017).

The most important benefit may come not in fundamental understanding but in engineering. The models in absolute units should better represent particular manipulations, such as altering the K_d for a particular clock protein binding to a particular promoter. This is the level of understanding that the Crops *in silico* initiative and others propose for some key processes in crop growth, in order to apply molecular genetic tools most powerfully to crop improvement (Zhu *et al.*, 2016). Detailed models will be required to design interventions in those processes, such as the comprehensive, OnGuard stomatal physiology model (Hills *et al.*, 2012) or the ePhotosynthesis model (Zhu *et al.*, 2013). The biochemical and biophysical parameter values in ePhotosynthesis derive from many species but none is from Arabidopsis. In part, this reflects the technical challenges that a very small plant presents for photosynthesis research (Stitt *et al.*, 2010). However, the (excellent) researcher who most directly measured parameter values for our clock models rated that as their most boring work ever, hinting at the social factors that also shape research.

Process and Pizzazz for a digital plant community

Crops *in silico* aims to link discovery science that is far from agricultural production, with crop models that are closely linked to practice (Figure 5). Such different research areas bring distinct types of social organisation, as Vermeulen pointed out in another context: "In (post-)genomics research understanding is geared towards innovation, which requires higher levels of integration [among research groups], while ecology research is primarily oriented towards understanding nature and environmental change, allowing more decoupled forms of organisation. This different orientation of molecular biology and ecology also causes a difference in financial resources for collaboration, as the goal of improving human health attracts more research funding than increased understanding of basic environmental processes." (Vermeulen *et al.*, 2013). The Crops *in silico* initiative foresees a substantial effort in social organisation, drawing from examples including SBML, the Physiome and "virtual organism" initiatives such as the Virtual Rat or

Virtual Physiological Human (Marshall-Colon *et al.*, 2017). These networked, interdisciplinary research organisations are an active domain for social science research, which is generating results and concepts that seem relevant for practitioners (Freeman and Millar, 2017). The "Community of Practice", for example, links members who share a common goal across the boundaries of previously-separate fields: Crops *in silico* seeks to establish such a community. One challenge is to attract members. The relative youth of the Arabidopsis field might offer some advantage here, in providing new members to an emerging plant modelling community (see final section).

The promise and challenge of shared resources

"Boundary Organisations" can also support the emerging community, particularly if they manage "Boundary Objects" (Star and Griesemer, 1989). These can be physical: the high-throughput plant phenotyping facilities and the EMPHASIS network that coordinates them in the EU form one example (Roy *et al.*, 2017). The Biomodels repository of models (Glont *et al.*, 2018) is such an Object from the Systems Biology community, and its original focus was on models in SBML format. Biomodels addresses a practical need specific to that community, attracts investments from different constituencies (models from biologists and software tools from computer science) and thereby creates a form of shared, social capital. Plant science is not, however, a major component: 38 models include Arabidopsis components or literature, of a total 1649 published models (in mid-2018); 2 models include maize references; 0 for wheat or barley. Biomodels policy is now to accept models in any format, increasing its relevance for crop models. It seems relevant that Biomodels is hosted by the European Bioinformatics Institute (EBI), itself part of the inter-governmental, treaty organisation EMBL (established 1974). One or more anchor institutions with stable mission and funding will be extremely beneficial for the risky, long-term development of complex plant models and their associated communities.

Crops *in silico* must link very diverse data with the diverse models, so resources to manage the data might form another, helpful, Boundary Object. Alongside the experimental phenotyping facilities mentioned above, data resources have been developed to manage and share large-scale plant phenotyping data (Neveu *et al.*, 2018). The Agricultural Models Intercomparison and Improvement Project (AgMIP) has worked to assemble benchmark data as well as crop models, for example (Asseng *et al.*, 2013; Rosenzweig *et al.*, 2013). Systems biology models, in contrast, are too rarely benchmarked: open, community-based benchmarking would help to give credit for model improvements. However, many of the data that we need are acquired at the single-project scale (as in Table 1), where data sharing is still not routine.

The Open Research movement (The Royal Society, 2012) promotes sharing of data (Open Data), as well as publications (Open Access), software (Open Source) and in some cases, even lab notebooks (Open Notebook Science). “Data” is very broadly conceived, including protocols, analysis or visualisation scripts, and models, as well as experimental data. The principles of FAIR data are more recent but equally important for Crops *in silico*, as they promote data that are Findable, Accessible, Interoperable and Re-usable (Wilkinson *et al.*, 2016). FAIR data need not be Open, but if access is granted then they should be easier to use. In contrast, Open data that is not FAIR might be unusable. FAIR is therefore being proposed as a guiding principle for international initiatives such as the European Open Science Cloud and the US NIH Data Commons (see Box 1).

To get FAIR data beyond the principles and into common research practice, we need easy-to-use software tools and resources. Resources to manage the “long-tail” data (Ferguson *et al.*, 2014) that is required for detailed modelling can in theory be “explicitly created to meet the researchers' needs, support extensive curation, and embody a heightened awareness of what it takes to make data re-useable by others” (Leonelli *et al.*, 2013). Although this is clearly desirable, few biology groups have such data management resources, or the software skills to customise them for their needs, or much appetite to add data curation to their overloaded schedules. The data curated in Table 1, for example, were assembled only because they were required for a specific research project. The software that might underpin such resources is fragmented (Kwok, 2018), except where research funders have coordinated internationally as in the AgMIP and FAIRDOM projects (see Box 1) (Rosenzweig *et al.*, 2013; Wolstencroft *et al.*, 2017). Coordination among funders, including direct funding for data curation, will be essential to get beyond pilot, example models and create a broadly-based digital organism framework that is regularly updated and refined with new information, in turn supporting the careers of a new generation of modellers.

Conclusion

No one should be surprised that such major research problems are relatively neglected, if funders, researchers and their institutions recognise and reward individual lab heads catching transient, project awards, like superhero characters in a video game. We have argued that projects should be valued, rather than individuals (Freeman and Millar, 2017). This requires the intellectual platform, capability and leadership to manage such projects, which is itself an area for rich debate (Mazzucato, 2014; Rip, 2000; Weber *et al.*, 2016). Large projects in this area require international, community-wide effort but this does not imply that they should be monolithic. Rather they need particular infrastructure, with funding mechanisms suited to infrastructure, to integrate the results from distributed projects that might be independently funded.

This article focussed on the need for digital organism initiatives to create and integrate a network of diverse models, and practical steps towards integration (summarised in Figure 5). Model diversity will always be with us, due to the variety of biological, chemical and physical processes involved, the uneven states of knowledge, mathematical and computational tools, and the differing aims of model users. Digital organism initiatives recognise both the model integration tasks and the parallel challenge of managing diverse data. We touched on the technical infrastructure that is required but community structures and community dynamics also contribute to the operation and governance of such research networks (Freeman and Millar, 2017). Social infrastructure therefore has a key role and might require parallel, infrastructural funding, which will change over time. Community organisation might initially focus on understanding and testing pilot model integrations, for example, whereas standardisation might be a later stage, as we noted in the case of SBML.

In a landscape of this complexity, engaging multiple research and stakeholder communities, projects like Crops *in silico* will be demanding of their leadership. The social sciences may contribute useful strategies (Balmer *et al.*, 2016) but these do little to mitigate the risks for junior faculty, until concerns over lower funding and recognition for interdisciplinary research are resolved (Bromham *et al.*, 2016; Rafols *et al.*, 2012; Yegros-Yegros *et al.*, 2015). We might rather harness the motivation of our youngest researchers. The success of the student-led International Genetically Engineered Machines competition (iGEM) brought a definite buzz to Synthetic Biology (Matheson, 2017), by giving them tools, keeping an open competition, and making it fun.

Acknowledgements

We gratefully acknowledge discussion with Prof. Vincent Danos (Informatics, University of Edinburgh) and Dr. Robert Muetzelfeldt (Simulistics Ltd., Edinburgh), Dr. Paul. E Brown (University of Warwick) for the simulations in Figure 3, along with funding from the UK Biotechnology and Biological Sciences Research Council (awards C006658, D019621, F010583, F005237, M017605, M018040) and the European Commission (FP7 TiMet, award 245143), and research studentships from BBSRC (iCASE K011294) and CONACYT (Mexico).

Text Box 1

Box 1: Online Resources and Software

- Agricultural Models Exchange Initiative (AMEI), repository of models and resources for model exchange in CropML, by Pierre Martre, Christophe Pradal *et al*. https://github.com/AgriculturalModelExchangeInitiative.
- Agricultural Models Intercomparison and Improvement Project (AgMIP), international programme of data format interconversion and model comparison for crop models, http://www.agmip.org.
- cis_interface, software tools to link "black box" models, by Meagan Lang (National Centre for Supercomputing Applications, Illinois, USA) https://github.com/cropsinsilico/cis_interface.
- European Open Science Cloud, high-level initiative in Open Research that includes FAIR data principles, https://ec.europa.eu/research/openscience/.
- FAIRDOM, international project developing software for "long-tail" research data management and advocating Open and FAIR data, https://fair-dom.org.
- FAIRDOMHub, instance of FAIRDOM software providing a self-service commons for public or private data, models and protocols, https://fairdomhub.org.
- GARNet (previously the Genomics Arabidopsis Research Network), organization representing the UK Arabidopsis research community; several relevant reports online: http://www.garnetcommunity.org.uk.
- NIH Data Commons, pilot project (2017-2020) including FAIR data principles, https://commonfund.nih.gov/commons.
- Plant Systems Modelling (PlaSMo), repository of plant growth models in several formats, https://www.plasmo.ed.ac.uk; now migrated to the FAIRDOMHub commons.
- SBMLDataTools, software tools to add external timeseries data as a function in an SBML model, by Alastair Hume (EPCC, Edinburgh, UK). https://github.com/allyhume/SBMLDataTools.
- SimileXMLv3, XML schema for Simile models, with a model conversion tool. http://www.simulistics.com/book/similexml/simile-markup-languages/similexmlv3 [the PlaSMo project presented a dozen models, refactored into this standard; Simile software support had lapsed at the time of writing].

Table 1. Parameter values for detailed modelling were collated from the literature.

[1] PMID, PubMed identifier of the publication.

					Publication reference				
Component	**Process**	**Sample**	**Value**	**units**	**PMID[1]**	**First Author**	**Year**	**Data display**	**Comments**
Cytosol	Volume	*A. thaliana* leaf	4.1	% of cell volume	23265941	Koffler BE	2013	Table 1	
Mitochondria	Volume	*A. thaliana* leaf	0.47	% of cell volume	23265941	Koffler BE	2013	Table 1	
Chloroplasts	Volume	*A. thaliana* leaf	15.63	% of cell volume	23265941	Koffler BE	2013	Table 1	
Nucleus	Volume	*A. thaliana* leaf	0.16	% of cell volume	23265941	Koffler BE	2013	Table 1	
Peroxisomes	Volume	*A. thaliana* leaf	0.14	% of cell volume	23265941	Koffler BE	2013	Table 1	
Vacuole	Volume	*A. thaliana* leaf	79.19	% of cell volume	23265941	Koffler BE	2013	Table 1	
Nucleus	Diameter	*A. thaliana* leaf	5.99	µm	19650905	Tirichine L	2009		
Cell	Volume	*A. thaliana* leaf	73000	µm^3	20598116	Wuyts N	2010	Fig. 8, left bottom	Mean value for palisade mesophyll cells.
Gene transcription	transcription rate	Yeast	2 - 30	mRNA/hour	21103382	Pelechano V	2010	Abstract	Reported range is 2-30 mRNA/hour.
RNA Polymerase II	density on DNA	Yeast	0.078	Pol II molecules/kb	21103382	Pelechano V	2010		
RNA Polymerase II	density on DNA	Yeast	2	pol II/gene	19011635	Zenklusen D	2008		
RNA Polymerase II	elongation rate	Yeast	0.56	kb/min	24103494	Miguel A	2013	Fig. 1A	21ºC
RNA Polymerase II	elongation rate	Mammalian cells	4	kb/min	21264352	Brody	2011		
RNA Polymerase II	elongation rate	Zebrafish	4.8	kb/min	23250218	Hanisch A	2013	Abstract	Measured at 28.5 ºC.
Ribosome density	Translation	*E. coli*	11 ± 2	ribosomes/RNA	19167328	Brandt F	2009	Fig. 2G	In polysomes translating firefly Luciferase
Nuclear Pore Complex (NPC)	density on nuclear envelope	lymphocytes	2 - 4	NPCs/µm^2	19392704	Fiserova	2009		
NPC	density on nuclear envelope	Mature Xenopus oocytes	60	NPCs/µm^2	19392704	Fiserova	2009		
NPC	density on nuclear envelope	Tobacco cell cultures	50	NPC/µm^2	19392704	Fiserova	2009		40-50 for 3-day-old cells; 50 for 10-day-old cells.
Transportin protein	Nuclear translocation rate	Mammalian (HeLa) cells	140	molecules/s/NPC	15657394	Kubitscheck U.	2005		
NTF2 protein	Nuclear translocation rate	Mammalian (HeLa) cells	170	molecules/s/NPC	15657394	Kubitscheck U.	2006		
Nucleoplasmin core domain fusion protein	Nuclear translocation rate	Mammalian (HeLa) cells	17	MDal/s/NPC	11250898	Ribbeck K.	2001		

Figure legends

Fig. 1. A model can usefully be represented in several forms.

(A) A simple model of the circadian clock gene circuit (Locke *et al.*, 2005) is shown as an informal diagram, linking four genes (helices) *via* their proteins (ovals), with inputs from light (sun). (B) The differential equation for changes in cytosolic LHY protein (cL_c) in the model is human-readable (and declarative). This equation involves *LHY* mRNA (cL_m), a translation rate parameter (p_1), RNA degradation rate parameters (m_2, k_2), and translocation of nuclear LHY protein (cL_n) with rates r_1, r_2. (C) A fragment of SBML represents the equation with the same names but is now machine-readable. The first line provides a stable reference to interpret its MathML format. (D) Timeseries simulation of the SBML model in suitable software provided a model output for the RNA level of gene *Y* (Y fit; red, open symbols; timepoints selected to match data), for comparison to RNA data acquired for a candidate gene in Arabidopsis (GI data, filled symbols). After a dark night (-12h to 0h), dawn light transiently induces both the hypothetical *Y* and candidate gene *GI;* the simulation continues in constant light. The comparison of model to data leads to future model refinement (dashed arrow) in the iterative cycle of systems biology. Adapted from (Locke *et al.*, 2005).

Fig. 2. The simple, qualitative form of a model can retain key behaviours.

(A) Simulation outputs show RNA levels changing continuously, from the simple clock model (Locke *et al.*, 2005) in quantitative form (differential equations, as in Figure 1B). (B) RNAs are either expressed (1) or not (0) in the qualitative form of the same model (Akman *et al.*, 2012). The binary, time-delay model still shows bimodal peaks of RNA expression from gene *Y* (green), with light induction after dawn (as in Figs. 1D, 2A). Levels are slightly offset for clarity in (B). Time 0h is midnight. Open box, light interval; filled box, dark interval.

Fig. 3. New capabilities arise from a “black-box” combination of models.

The circadian clock model shown in Figure 1 (Locke *et al.*, 2005) can communicate to the Arabidopsis architectural model (Mundermann *et al.*, 2005) running in L-studio software. A version of the clock model in Matlab software was automatically compiled into the C programming language (creating a ‘black box’), in order to interact as a black box with the lpfg programme of L-studio. TOC1 protein level from the clock model controlled a leaf angle parameter in the architectural model, creating a simple simulation of rhythmic leaf movement in Arabidopsis over day/night cycle. The clock model’s light:dark setting also darkened plant colour at night (16h, 20h). Image generated by Paul E. Brown and A.J. Millar.

Fig. 4. Published parameter values can inform detailed modelling.

(A) Distribution of published K_d values for plant protein-protein interaction affinities. (B) Distribution of published K_d values for plant protein-protein interaction affinities. In the (many) cases where an interaction of interest has not been measured directly, data such as these help to constrain the range of parameter values that computational, parameter-fitting procedures should explore. Publication references are listed in the main text.

Fig. 5. Linking Systems Biology with Crop Science models.

The solid line links the concepts of biology, first from genome sequence *via* genotype, biochemical parameters and molecular regulation to whole-organism phenotype in a particular environment (yellow area); then from phenotypes to field traits and adaptation or to yield under particular management (green area); finally, given genetic variation, through natural selection or artificial selection in crop breeding, to the evolution of genome sequences (adapted from Millar, 2016). Initiatives like Crops *in silico* will deal with the whole cycle, by linking several models (coloured arcs) into a seamless, causal chain. The top line of graphics locate the topics considered in the main text with reference to this cycle. The arcs suggest current types of model, in systems biology (indigo), crop science (cyan) and evolution (dark blue). The dimensions that are often considered in such models are capitalized (G, P, E, M). Underpinning infrastructures (grey) help to bridge these disciplines. ‘Anchor’ institutions are shown (buildings), which might provide major experimental facilities, digital infrastructure or a focus for social infrastructure, such as training or standardisation workshops.

References

Adam M, Corbeels M, Leffelaar PA, Van Keulen H, Wery J, Ewert F. 2012. Building crop models within different crop modelling frameworks. Agricultural Systems **113**, 57-63.

Adams RR, Tsorman N, Stratford K, Akman OE, Gilmore S, Juty N, Le Novere N, Millar AJ. 2012. The Input Signal Step Function (ISSF), a standard method to encode input signals in SBML models with software support, applied to circadian clock models. Journal of Biological Rhythms **27**, 328-332.

Aderhold A. HD, Smith V.A., Millar A.J., and Grzegorczyk M. 2013. Assessment of Regression Methods for inference of regulatory networks involved in circadian regulation. *WCSB2013, 10th International Workshop on Computational Systems Biology*. Tampere, Finland 29-33.

Aggarwal P, Das Gupta M, Joseph AP, Chatterjee N, Srinivasan N, Nath U. 2010. Identification of specific DNA binding residues in the TCP family of transcription factors in Arabidopsis. The Plant Cell **22**, 1174-1189.

Akman OE, Watterson S, Parton A, Binns N, Millar AJ, Ghazal P. 2012. Digital clocks: simple Boolean models can quantitatively describe circadian systems. Journal of the Royal Society Interface **9**, 2365-2382.

Alberghina L, Westerhoff HV, eds. 2005. *Systems Biology, Definitions and Perspectives*: Springer-Verlag.

Alves R, Antunes F, Salvador A. 2006. Tools for kinetic modeling of biochemical networks. Nature Biotechnology **24**, 667-672.

Ankeny RA, Leonelli S. 2011. What's so special about model organisms? Studies in History and Philosophy of Science **42**, 313-323.

Asseng S, Ewert F, Rosenzweig C, *et al.* 2013. Uncertainty in simulating wheat yields under climate change. Nature Climate Change **3**, 827-832.

Baker RE, Pena JM, Jayamohan J, Jerusalem A. 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biology Letters **14**.

Ballut L, Drucker M, Pugniere M, *et al.* 2005. HcPro, a multifunctional protein encoded by a plant RNA virus, targets the 20S proteasome and affects its enzymic activities. Journal of General Virology **86**, 2595-2603.

Balmer AS, Calvert J, Marris C, Molyneux-Hodgson S, Frow E, Kearnes M, Bulpin K, Schyfter P, Mackenzie A, Martin P. 2016. Five rules of thumb for post-ELSI interdisciplinary collaborations. Journal of Responsible Innovation **3**, 73-80.

Bard JBL. 2008. Waddington’s Legacy to Developmental and Theoretical Biology. Biological Theory **3**, 188-197.

Baudry A, Ito S, Song YH, *et al.* 2010. F-box proteins FKF1 and LKP2 act in concert with ZEITLUPE to control Arabidopsis clock progression. The Plant Cell **22**, 606-622.

Bauer J, Reiss K, Veerabagu M, Heunemann M, Harter K, Stehle T. 2013. Structure-function analysis of Arabidopsis thaliana histidine kinase AHK5 bound to its cognate phosphotransfer protein AHP1. Molecular Plant **6**, 959-970.

Bergmann FT, Adams R, Moodie S, *et al.* 2014. COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. BMC Bioinformatics **15**, 369.

Bernal-Bayard P, Ojeda V, Hervás M, Cejudo FJ, Navarro JA, Velázquez-Campoy A, Pérez-Ruiz JM. 2014. Molecular recognition in the interaction of chloroplast 2-Cys peroxiredoxin with NADPH-thioredoxin reductase C (NTRC) and thioredoxinx. FEBS Letters **588**, 4342-4347.

Bisson MM, Groth G. 2010. New insight in ethylene signaling: autokinase activity of ETR1 modulates the interaction of receptors and EIN2. Molecular Plant **3**, 882-889.

Bothwell JHF. 2006. The long past of systems biology. New Phytologist **170**, 6-10.

Brandt F, Etchells SA, Ortiz JO, Elcock AH, Hartl FU, Baumeister W. 2009. The native 3D organization of bacterial polysomes. Cell **136**, 261-271.

Bromham L, Dinnage R, Hua X. 2016. Interdisciplinary research has consistently lower funding success. Nature **534**, 684-687.

Buchel F, Rodriguez N, Swainston N, *et al.* 2013. Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Systems Biology **7**, 116.

Bujdoso N, Davis SJ. 2013. Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of Arabidopsis thaliana. Frontiers in Plant Science **4**, 3.

Caron E, Ghosh S, Matsuoka Y, Ashton-Beaucage D, Therrien M, Lemieux S, Perreault C, Roux PP, Kitano H. 2010. A comprehensive map of the mTOR signaling network. Molecular Systems Biology **6**, 453.

Cheung CY, Poolman MG, Fell DA, Ratcliffe RG, Sweetlove LJ. 2014. A Diel Flux Balance Model Captures Interactions between Light and Dark Metabolism during Day-Night Cycles in C3 and Crassulacean Acid Metabolism Leaves. Plant Physiology **165**, 917-929.

Chew YH, Seaton DD, Mengin V, Flis A, Mugford ST, Smith AM, Stitt M, Millar AJ. 2017. Linking circadian time to growth rate quantitatively via carbon metabolism. bioRxiv 10.1101/105437.

Chew YH, Smith RW, Jones HJ, Seaton DD, Grima R, Halliday KJ. 2014a. Mathematical models light up plant signaling. The Plant Cell **26**, 5-20.

Chew YH, Wenden B, Flis A, *et al.* 2014b. Multiscale digital Arabidopsis predicts individual organ and whole-organism growth. Proceedings of the National Academy of Sciences of the USA **111**, E4127-4136.

Coveney PV, Fowler PW. 2005. Modelling biological complexity: a physical scientist's perspective. Journal of the Royal Society Interface **2**, 267-280.

Dalchau N, Baek SJ, Briggs HM, *et al.* 2011. The circadian oscillator gene GIGANTEA mediates a long-term response of the Arabidopsis thaliana circadian clock to sucrose. Proceedings of the National Academy of Sciences of the U S A **108**, 5104-5109.

Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, Siepel A, Kraus WL. 2013. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Molecular Cell **50**, 212-222.

Davey C, Ougham H, Millar A, Thomas H, Tindal C, Muetzelfeldt R. 2009. PlaSMo: Making existing plant and crop mathematical models available to plant systems biologists. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology **153**, S225-S226.

De Jong H. 2002. Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology **9**, 67-103.

Dempsey L. 2013. The Inside-Out Library. *On libraries, services and networks*: Slideshare.

Diaz S, Kattge J, Cornelissen JH, *et al.* 2016. The global spectrum of plant form and function. Nature **529**, 167-171.

Dong CH, Jang M, Scharein B, Malach A, Rivarola M, Liesch J, Groth G, Hwang I, Chang C. 2010. Molecular association of the Arabidopsis ETR1 ethylene receptor and a regulator of ethylene signaling, RTE1. Journal of Biological Chemistry **285**, 40706-40713.

Dyson RJ, Vizcay-Barrena G, Band LR, *et al.* 2014. Mechanical modelling quantifies the functional importance of outer tissue layers during root elongation and bending. New Phytologist **202**, 1212-1222.

Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER. 2004. A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. The Plant Cell **16**, 2923-2939.

Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE, Martone ME. 2014. Big data from small data: data-sharing in the 'long tail' of neuroscience. Nature Neuroscience **17**, 1442-1447.

Finkenstadt B, Heron EA, Komorowski M, Edwards K, Tang S, Harper CV, Davis JR, White MR, Millar AJ, Rand DA. 2008. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. Bioinformatics **24**, 2901-2907.

Fiserova J, Kiseleva E, Goldberg MW. 2009. Nuclear envelope and nuclear pore complex structure and organization in tobacco BY-2 cells. Plant Journal **59**, 243-255.

Flis A, Fernandez AP, Zielinski T, *et al.* 2015. Defining the robust behaviour of the plant clock gene circuit with absolute RNA timeseries and open infrastructure. Open Biology **5**, 150042.

Flis A, Mengin V, Ivakov AA, *et al.* 2018. Multiple circadian clock outputs regulate diel turnover of carbon and nitrogen reserves. Plant, Cell & Environment **in press**.

Fogelmark K, Troein C. 2014. Rethinking transcriptional activation in the Arabidopsis circadian clock. PLoS Computational Biology **10**, e1003705.

Freeman PL, Millar AJ. 2017. Valuing the project: a knowledge-action response to network governance in collaborative research. Public Money & Management **37**, 23-30.

Fuglsang AT, Borch J, Bych K, Jahn TP, Roepstorff P, Palmgren MG. 2003. The binding site for regulatory 14-3-3 protein in plant plasma membrane H+-ATPase: involvement of a region promoting phosphorylation-independent interaction in addition to the phosphorylation-dependent C-terminal end. Journal of Biological Chemistry **278**, 42266-42272.

Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H. 2008. CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. Proceedings of the IEEE **96**, 1254-1265.

GARNet Advisory Committee. 2006. Final report of the GARNet Advisory Committee on Arabidopsis Systems Biology in the UK, June 2006. Millar AJ, ed.

Ghosh S, Matsuoka Y, Asai Y, Hsin K-Y, Kitano H. 2011. Software for systems biology: from tools to integrated platforms. Nature Reviews Genetics **12**, 821-832.

Glont M, Nguyen TVN, Graesslin M, *et al.* 2018. BioModels: expanding horizons to include more modelling approaches and formats. Nucleic Acids Research **46**, D1248-D1253.

Goldberg AP, Szigeti B, Chew YH, Sekar JA, Roth YD, Karr JR. 2018. Emerging whole-cell modeling principles and methods. Current Opinion in Biotechnology **51**, 97-102.

Goncalves E, van Iersel M, Saez-Rodriguez J. 2013. CySBGN: a Cytoscape plug-in to integrate SBGN maps. BMC Bioinformatics **14**, 17.

Gould PD, Domijan M, Greenwood M, Tokuda IT, Rees H, Kozma-Bognar L, Hall AJ, Locke JC. 2018. Coordination of robust single cell rhythms in the Arabidopsis circadian clock via spatial waves of gene expression. Elife **7**, 31700.

Grzegorczyk M, Aderhold A, Husmeier D. 2015. Inferring bi-directional interactions between circadian clock genes and metabolism with model ensembles. Statistical Applications in Genetics and Molecular Biology **2014**, 0041.

Grzegorczyk M, Husmeier D, Edwards KD, Ghazal P, Millar AJ. 2008. Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. Bioinformatics **24**, 2071-2078.

Guerriero ML, Pokhilko A, Fernandez AP, Halliday KJ, Millar AJ, Hillston J. 2012. Stochastic properties of the plant circadian clock. Journal of the Royal Society Interface **9**, 744-756.

Hammer GL, Sinclair TR, Chapman SC, van Oosterom E. 2004. On systems thinking, systems biology, and the in silico plant. Plant Physiology **134**, 909-911.

Hanisch A, Holder MV, Choorapoikayil S, Gajewski M, Ozbudak EM, Lewis J. 2013. The elongation rate of RNA polymerase II in zebrafish and its significance in the somite segmentation clock. Development **140**, 444-453.

Hao D, Ohme-Takagi M, Sarai A. 1998. Unique mode of GCC box recognition by the DNA-binding domain of ethylene-responsive element-binding factor (ERF domain) in plant. Journal of Biological Chemistry **273**, 26857-26861.

Hao Q, Yin P, Li W, Wang L, Yan C, Lin Z, Wu Jim Z, Wang J, Yan SF, Yan N. 2011. The Molecular Basis of ABA-Independent Inhibition of PP2Cs by a Subclass of PYL Proteins. Molecular Cell **42**, 662-672.

Harper CV, Finkenstadt B, Woodcock DJ, *et al.* 2011. Dynamic analysis of stochastic transcription cycles. PLoS Biology **9**, e1000607.

Henry J. 2008. Historical and other studies of science, technology and medicine in the University of Edinburgh. Notes and Records of the Royal Society **62**, 223-235.

Higham CF, Husmeier D. 2013. A Bayesian approach for parameter estimation in the extended clock gene circuit of Arabidopsis thaliana. BMC Bioinformatics **14**, S3.

Hills A, Chen ZH, Amtmann A, Blatt MR, Lew VL. 2012. OnGuard, a computational platform for quantitative kinetic modeling of guard cell physiology. Plant Physiology **159**, 1026-1042.

Hofr C, Sultesova P, Zimmermann M, Mozgova I, Prochazkova Schrumpfova P, Wimmerova M, Fajkus J. 2009. Single-Myb-histone proteins from Arabidopsis thaliana: a quantitative study of telomere-binding specificity and kinetics. Biochemical Journal **419**, 221-228.

Honorato-Zimmer R, Millar AJ, Plotkin GD, Zardilis A. 2017. Chromar, a language of parameterised objects. Theoretical Computer Science **7**, 34.

Hucka M, Bergmann FT, Drager A, *et al.* 2018. The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core. Journal of Integrative Bioinformatics **2017**, 081.

Hucka M, Finney A, Sauro HM, *et al.* 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19**, 524-531.

Hucka M, Nickerson DP, Bader GD, *et al.* 2015. Promoting Coordinated Development of Community-Based Information Standards for Modeling in Biology: The COMBINE Initiative. Frontiers in Bioengineering and Biotechnology **3**, 19.

Ishihara H, Obata T, Sulpice R, Fernie AR, Stitt M. 2015. Quantifying protein synthesis and degradation in Arabidopsis by dynamic 13CO2 labeling and analysis of enrichment in individual amino acids in their free pools and in protein. Plant Physiology **168**, 74-93.

Izawa T, Foster R, Chua N-H. 1993. Plant bZIP protein DNA binding specificity. Journal of Molecular Biology **230**, 1131-1144.

Jagaman K, Danuser G. 2006. Linking data to models: data regression. Nature Reviews Molecular Cell Biology **7**, 813-819.

Janes KA, Yaffe MB. 2006. Data-driven modelling of signal-transduction networks. Nature Reviews Molecular Cell Biology **7**, 820-828.

Jonsson H, Heisler MG, Shapiro BE, Meyerowitz EM, Mjolsness E. 2006. An auxin-driven polarized transport model for phyllotaxis. Proceedings of the National Academy of Sciences of the U S A **103**, 1633-1638.

Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Jr., Assad-Garcia N, Glass JI, Covert MW. 2012. A whole-cell computational model predicts phenotype from genotype. Cell **150**, 389-401.

Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J. 2011. A structure-based benchmark for protein-protein binding affinity. Protein Science **20**, 482-491.

Kierzkowski D, Nakayama N, Routier-Kierzkowska AL, Weber A, Bayer E, Schorderet M, Reinhardt D, Kuhlemeier C, Smith RS. 2012. Elastic domains regulate growth and organogenesis in the plant shoot apical meristem. Science **335**, 1096-1099.

Kitano H. 2002. Systems biology: a brief overview. Science **295**, 1662-1664.

Kitano H. 2015. Accelerating systems biology research and its real world deployment. NPJ Systems Biology and Applications **1**, 15009.

Kniemeyer O, Buck-Sorlin G, Kurth W. 2007. GroIMP as a platform for functional-structural modelling of plants. Functional-Structural Plant Modelling in Crop Production **22**, 43-52.

Knowles SM, Lu SX, Tobin EM. 2008. Testing time: can ethanol-induced pulses of proposed oscillator components phase shift rhythms in Arabidopsis? Journal of Biological Rhythms **23**, 463-471.

Koffler BE, Bloem E, Zellnig G, Zechmann B. 2013. High resolution imaging of subcellular glutathione concentrations by quantitative immunoelectron microscopy in different leaf areas of Arabidopsis. Micron **45**, 119-128.

Kubitscheck U, Grunwald D, Hoekstra A, Rohleder D, Kues T, Siebrasse JP, Peters R. 2005. Nuclear transport of single molecules: dwell times at the nuclear pore complex. Journal of Cell Biology **168**, 233-243.

Kumar MD, Gromiha MM. 2006. PINT: Protein-protein Interactions Thermodynamic Database. Nucleic Acids Research **34**, D195-198.

Kwok R. 2018. How to pick an electronic laboratory notebook. Nature **560**, 269-270.

Le Novere N. 2015. Quantitative and logic modelling of molecular and gene networks. Nature Reviews Genetics **16**, 146-158.

Le Novere N, Hucka M, Mi H, *et al.* 2009. The Systems Biology Graphical Notation. Nature Biotechnology **27**, 735-741.

Leitner F, Bielza C, Hill SL, Larranaga P. 2016. Data Publications Correlate with Citation Impact. Frontiers in Neuroscience **10**, 419.

Leonelli S. 2007. Growing weed, producing knowledge an epistemic history of Arabidopsis thaliana. History and Philosophy of the Life Sciences **29**, 193-223.

Leonelli S, Smirnoff N, Moore J, Cook C, Bastow R. 2013. Making open data work for plant scientists. Journal of Experimental Botany **64**, 4109-4117.

Levskaya A, Weiner OD, Lim WA, Voigt CA. 2009. Spatiotemporal control of cell signalling using a light-switchable protein interaction. Nature **461**, 997-1001.

Li J, Smith GP, Walker JC. 1999. Kinase interaction domain of kinase-associated protein phosphatase, a phosphoprotein-binding domain. Proceedings of the National Academy of Sciences of the USA **96**, 7821-7826.

Li L, Nelson CJ, Trosch J, Castleden I, Huang S, Millar AH. 2017. Protein Degradation Rate in Arabidopsis thaliana Leaf Growth and Development. The Plant Cell **29**, 207-228.

Liang X, Nazarenus TJ, Stone JM. 2008. Identification of a consensus DNA-binding site for the Arabidopsis thaliana SBP domain transcription factor, AtSPL14, and binding kinetics by surface plasmon resonance. Biochemistry **47**, 3645-3653.

Libault M, Pingault L, Zogli P, Schiefelbein J. 2017. Plant Systems Biology at the Single-Cell Level. Trends in Plant Science **22**, 949-960.

Lidder P, Gutierrez RA, Salome PA, McClung CR, Green PJ. 2005. Circadian control of messenger RNA stability. Association with a sequence-specific messenger RNA decay pathway. Plant Physiology **138**, 2374-2385.

Liu X, Yue Y, Li B, Nie Y, Li W, Wu WH, Ma L. 2007. A G protein-coupled receptor is a plasma membrane receptor for the plant hormone abscisic acid. Science **315**, 1712-1716.

Lloyd CM, Halstead MD, Nielsen PF. 2004. CellML: its future, present and past. Progress in Biophysics and Molecular Biology **85**, 433-450.

Locke JC, Southern MM, Kozma-Bognar L, Hibberd V, Brown PE, Turner MS, Millar AJ. 2005. Extension of a genetic network model by iterative experimentation and mathematical analysis. Molecular Systems Biology **1**, 2005 0013.

Luoni L, Bonza MC, De Michelis MI. 2006. Calmodulin/Ca2+-ATPase interaction at the Arabidopsis thaliana plasma membrane is dependent on calmodulin isoform showing isoform-specific Ca2+ dependencies. Physiologia Plantarum **126**, 175-186.

Ma C, Zhang HH, Wang X. 2014. Machine learning for Big Data analytics in plants. Trends in Plant Science **19**, 798-808.

Macklin DN, Ruggero NA, Covert MW. 2014. The future of whole-cell modeling. Current Opinion in Biotechnology **28C**, 111-115.

Mantovani R, Aguilar X, Blomberg J, Brännström K, Olofsson A, Schleucher J, Björklund S. 2014. Interaction Studies of the Human and Arabidopsis thaliana Med25-ACID Proteins with the Herpes Simplex Virus VP16- and Plant-Specific Dreb2a Transcription Factors. PLoS ONE **9**, e98575.

Marcum JA. 2008. Does systems biology represent a Kuhnian paradigm shift? New Phytologist **179**, 587-589.

Marji M. 2014. *Learn to program with Scratch : a visual introduction to programming with games, art, science, and math*. San Francisco: No Starch Press.

Marshall-Colon A, Long SP, Allen DK, *et al.* 2017. Crops In Silico: Generating Virtual Crops Using an Integrative and Multi-scale Modeling Platform. Frontiers in Plant Science **8**, 786.

Martre P, Donatelli M, Pradal C, *et al.* 2018. The Agricultural Model Exchange Initiative. *7th AgMIP Global Workshop*. IICA, San José, Costa Rica: IICA.

Matheson S. 2017. Engineering a Biological Revolution. Cell **168**, 329-332.

Mazzucato M. 2014. The entrepreneurial state : debunking public vs. private sector myths. London ; New York: Anthem Press.

Mendes P. 2018. Reproducible Research Using Biomodels. Bulletin of Mathematical Biology **80**, 3081-3087.

Middleton AM, Farcot E, Owen MR, Vernoux T. 2012. Modeling regulatory networks to understand plant development: small is beautiful. The Plant Cell **24**, 3876-3891.

Millar AJ. 2016. The intracellular dynamics of circadian clocks reach for the light of ecology and evolution Annual Review of Plant Biology **67**, 595-618.

Milo R, Jorgensen P, Moran U, Weber G, Springer M. 2010. BioNumbers--the database of key numbers in molecular and cell biology. Nucleic Acids Research **38**, D750-753.

Mjolsness E. 2018. Prospects for Declarative Mathematical Modeling of Complex Biological Systems. arXiv **1804**, 11044.

Moyroud E, Reymond MC, Hames C, Parcy F, Scutt CP. 2009. The analysis of entire gene promoters by surface plasmon resonance. Plant Journal **59**, 851-858.

Muetzelfeldt R. 2007. Declarative modelling in the ecological and environmental sciences. Nature Precedings **2007**, 17.

Muetzelfeldt R. 2010. A Unified Approach for Representing Structurally-Complex Models in SBML Level 3. Nature Precedings **2010**, 4372.

Muetzelfeldt R, Massheder J. 2003. The Simile visual modelling environment. European Journal of Agronomy **18**, 345-358.

Mundermann L, Erasmus Y, Lane B, Coen E, Prusinkiewicz P. 2005. Quantitative modeling of Arabidopsis development. Plant Physiology **139**, 960-968.

Naithani S, Preece J, D'Eustachio P, *et al.* 2017. Plant Reactome: a resource for plant pathways and comparative analysis. Nucleic Acids Research **45**, D1029-D1039.

Nakamichi N, Kiba T, Henriques R, Mizuno T, Chua NH, Sakakibara H. 2010. PSEUDO-RESPONSE REGULATORS 9, 7, and 5 are transcriptional repressors in the Arabidopsis circadian clock. The Plant Cell **22**, 594-605.

Narsai R, Howell KA, Millar AH, O'Toole N, Small I, Whelan J. 2007. Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. The Plant Cell **19**, 3418-3436.

Ndour A, Vadez V, Pradal C, Lucas M. 2017. Virtual Plants Need Water Too: Functional-Structural Root System Models in the Context of Drought Tolerance Breeding. Frontiers in Plant Science **8**, 1577.

Neveu P, Tireau A, Hilgert N, *et al.* 2018. Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. New Phytologist **in press**.

Novak B, Tyson JJ. 2008. Design principles of biochemical oscillators. Nature Reviews Molecular Cell Biology **9**, 981-991.

O'Neill JS, van Ooijen G, Le Bihan T, Millar AJ. 2011. Circadian clock parameter measurement: characterization of clock transcription factors using surface plasmon resonance. Journal of Biological Rhythms **26**, 91-98.

Ocone A, Millar AJ, Sanguinetti G. 2013. Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics. Bioinformatics **29**, 910-916.

Ogawa M, Shinohara H, Sakagami Y, Matsubayashi Y. 2008. Arabidopsis CLV3 peptide directly binds CLV1 ectodomain. Science **319**, 294.

Onoda Y, Wright IJ, Evans JR, Hikosaka K, Kitajima K, Niinemets U, Poorter H, Tosens T, Westoby M. 2017. Physiological and structural tradeoffs underlying the leaf economics spectrum. New Phytologist **214**, 1447-1463.

Ortiz-Gutierrez E, Garcia-Cruz K, Azpeitia E, Castillo A, Sanchez Mde L, Alvarez-Buylla ER. 2015. A Dynamic Gene Regulatory Network Model That Recovers the Cyclic Behavior of Arabidopsis thaliana Cell Cycle. PLoS Computational Biology **11**, e1004486.

Piques M, Schulze WX, Hohne M, Usadel B, Gibon Y, Rohwer J, Stitt M. 2009. Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. Molecular Systems Biology **5**, 314.

Pokhilko A, Fernandez AP, Edwards KD, Southern MM, Halliday KJ, Millar AJ. 2012. The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. Molecular Systems Biology **8**, 574.

Pokhilko A, Mas P, Millar AJ. 2013. Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs. BMC Systems Biology **7**, 23.

Poorter H, Niinemets U, Walter A, Fiorani F, Schurr U. 2010. A method to construct dose-response curves for a wide range of environmental factors and plant traits by means of a meta-analysis of phenotypic data. Journal of Experimental Botany **61**, 2043-2055.

Poorter H, Niklas KJ, Reich PB, Oleksyn J, Poot P, Mommer L. 2012. Biomass allocation to leaves, stems and roots: meta-analyses of interspecific variation and environmental control. New Phytologist **193**, 30-50.

Pradal C, Dufour-Kowalski S, Boudon F, Fournier C, Godin C. 2008. OpenAlea: a visual programming and component-based software platform for plant modelling. Functional Plant Biology **35**, 751.

Prouse MB, Campbell MM. 2013. Interactions between the R2R3-MYB transcription factor, AtMYB61, and target DNA binding sites. PLoS ONE **8**, e65132.

Prusinkiewicz P, Runions A. 2012. Computational models of plant development and form. New Phytologist **193**, 549-569.

Pudasaini A, Shim JS, Song YH, Shi H, Kiba T, Somers DE, Imaizumi T, Zoltowski BD. 2017. Kinetics of the LOV domain of ZEITLUPE determine its circadian function in Arabidopsis. Elife **6**, e21646.

Rafols I, Leydesdorff L, O'Hare A, Nightingale P, Stirling A. 2012. How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management. Research Policy **41**, 1262-1282.

Rand DA, Shulgin BV, Salazar JD, Millar AJ. 2006. Uncovering the design principles of circadian clocks: mathematical analysis of flexibility and evolutionary goals. Journal of Theoretical Biology **238**, 616-635.

Reuille PBd, Bohn-Courseau I, Ljung K, Morin h, Carraro N, Godin C, Traas J. 2006. Computer simulations reveal properties of the cell-cell signaling network at the shoot apex in Arabidopsis. Proceedings of the National Academy of Sciences of the USA **103**, 1627-1632.

Reymond MC, Brunoud G, Chauvet A, Martinez-Garcia JF, Martin-Magniette ML, Moneger F, Scutt CP. 2012. A Light-Regulated Genetic Module Was Recruited to Carpel Development in Arabidopsis following a Structural Change to SPATULA. The Plant cell **24**, 2812-2825.

Rip A. 2000. Higher forms of nonsense. European Review **8**, 467-485.

Rohn H, Junker A, Hartmann A, Grafahrend-Belau E, Treutler H, Klapperstuck M, Czauderna T, Klukas C, Schreiber F. 2012. VANTED v2: a framework for systems biology applications. BMC Systems Biology **6**, 139.

Rosenzweig C, Jones JW, Hatfield JL, *et al.* 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. Agricultural and Forest Meteorology **170**, 166-182.

Roy J, Tardieu F, Tixier-Boichard M, Schurr U. 2017. European infrastructures for sustainable agriculture. Nature Plants **3**, 756-758.

Schloss PD. 2017. Preprinting Microbiology. MBio **8**.

Schreiber F, Bader GD, Golebiewski M, *et al.* 2015. Specifications of Standards in Systems and Synthetic Biology. Journal of Integrative Bioinformatics **12**, 258.

Shahrezaei V, Swain PS. 2008. The stochastic nature of biochemical networks. Current Opinion in Biotechnology **19**, 369-374.

Sidaway-Lee K, Costa MJ, Rand DA, Finkenstadt B, Penfield S. 2014. Direct measurement of transcription rates reveals multiple mechanisms for configuration of the Arabidopsis ambient temperature response. Genome Biology **15**, R45.

SMBL community. 2017. SBML Specification Documents. Vol. 2018.

Snoep JL, Olivier BG. 2002. Java Web Simulation (JWS); a web based database of kinetic models. Molecular Biology Reports **29**, 259-263.

Sorokina O, Corellou F, Dauvillee D, Sorokin A, Goryanin I, Ball S, Bouget FY, Millar AJ. 2011. Microarray data can predict diurnal changes of starch content in the picoalga Ostreococcus. BMC Systems Biology **5**, 36.

Star SL, Griesemer JR. 1989. Institutional Ecology, Translations and Boundary Objects - Amateurs and Professionals in Berkeleys-Museum-of-Vertebrate-Zoology, 1907-39. Social Studies of Science **19**, 387-420.

Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, Allen JE, Zaslavskaia LA, MacBeath G. 2007. PDZ domain binding selectivity is optimized across the mouse proteome. Science **317**, 364-369.

Stitt M, Lunn J, Usadel B. 2010. Arabidopsis and primary photosynthetic metabolism - more than the icing on the cake. Plant Journal **61**, 1067-1091.

The Royal Society. (29 June 2012) Science as an open enterprise.

Thomas H. 2007. Systems biology and the biology of systems: how, if at all, are they related? New Phytologist **177**, 11-15.

Tirichine L, Andrey P, Biot E, Maurin Y, Gaudin V. 2009. 3D fluorescent in situ hybridization using Arabidopsis leaf cryosections and isolated nuclei. Plant Methods **5**, 11.

Truskina J, Vernoux T. 2018. The growth of a stable stationary structure: coordinating cell behavior and patterning at the shoot apical meristem. Current Opinion in Plant Biology **41**, 83-88.

Tyson JJ, Novak B. 2015. Models in biology: lessons from modeling regulation of the eukaryotic cell cycle. BMC Biology **13**, 46.

Valladares F, Matesanz S, Guilhaumon F, *et al.* 2014. The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. Ecology Letters **17**, 1351-1364.

van Ittersum MK, Ewert F, Heckelei T, *et al.* 2008. Integrated assessment of agricultural systems – A component-based framework for the European Union (SEAMLESS). Agricultural Systems **96**, 150-165.

Vermeulen N. 2017. The choreography of a new research field: Aggregation, circulation and oscillation. Environment and Planning A, 0308518X1772531.

Vermeulen N, Parker JN, Penders B. 2013. Understanding life together: a brief history of collaboration in biology. Endeavour **37**, 162-171.

Waltemath D, Adams R, Bergmann FT, *et al.* 2011. Reproducible computational biology experiments with SED-ML--the Simulation Experiment Description Markup Language. BMC Systems Biology **5**, 198.

Waltemath D, Karr JR, Bergmann FT, *et al.* 2016. Toward Community Standards and Software for Whole-Cell Modeling. IEEE Trans Biomed Eng **63**, 2007-2014.

Weber KM, Amanatidou E, Erdmann L, Nieminen M. 2016. Research and innovation futures: exploring new ways of doing and organizing knowledge creation. Foresight **18**, 193-203.

Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 160018.

Wittmann DM, Krumsiek J, Saez-Rodriguez J, Lauffenburger DA, Klamt S, Theis FJ. 2009. Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. BMC Systems Biology **3**, 98.

Wolstencroft K, Krebs O, Snoep JL, *et al.* 2017. FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. Nucleic Acids Research **45**, D404-D407.

Wolstencroft K, Owen S, Krebs O, *et al.* 2015. SEEK: a systems biology data and model management platform. BMC Systems Biology **9**, 33.

Wuyts N, Palauqui JC, Conejero G, Verdeil JL, Granier C, Massonnet C. 2010. High-contrast three-dimensional imaging of the Arabidopsis leaf enables the analysis of cell dimensions in the epidermis and mesophyll. Plant Methods **6**, 17.

Xuan W, Band LR, Kumpf RP, *et al.* 2016. Cyclic programmed cell death stimulates hormone signaling and root development in Arabidopsis. Science **351**, 384-387.

Yegros-Yegros A, Rafols I, D'Este P. 2015. Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity. PLoS ONE **10**, e0135095.

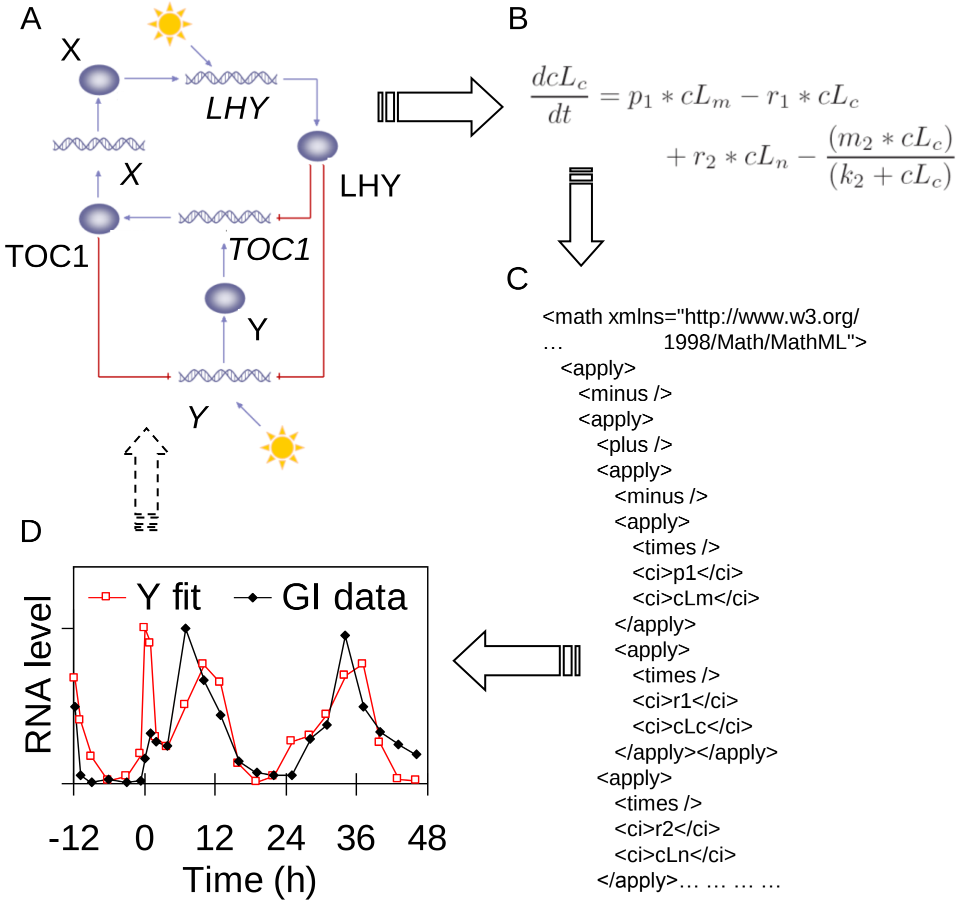
Zardilis A, Hume A, Millar AJ. 2019. A multi-model framework for the Arabidopsis life cycle. Journal of Experimental Botany **in press**.

Zenklusen D, Larson DR, Singer RH. 2008. Single-RNA counting reveals alternative modes of gene expression in yeast. Nature Structural & Molecular Biology **15**, 1263-1271.

Zhu XG, Lynch JP, LeBauer DS, Millar AJ, Stitt M, Long SP. 2016. Plants in silico: why, why now and what?--an integrative platform for plant systems biology research. Plant Cell and Environment **39**, 1049-1057.

Zhu XG, Wang Y, Ort DR, Long SP. 2013. e-Photosynthesis: a comprehensive dynamic mechanistic model of C3 photosynthesis: from light capture to sucrose synthesis. Plant Cell and Environment **36**, 1711-1727.

Figure 1



$$\frac{dcL_c}{dt} = p_1 * cL_m - r_1 * cL_c + r_2 * cL_n - \frac{(m_2 * cL_c)}{(k_2 + cL_c)}$$

```
<math xmlns="http://www.w3.org/
…             1998/Math/MathML">
  <apply>
    <minus />
    <apply>
      <plus />
      <apply>
        <minus />
        <apply>
          <times />
          <ci>p1</ci>
          <ci>cLm</ci>
        </apply>
        <apply>
          <times />
          <ci>r1</ci>
          <ci>cLc</ci>
        </apply></apply>
      <apply>
        <times />
        <ci>r2</ci>
        <ci>cLn</ci>
      </apply>… … … …
```

Fig. 1. A model can usefully be represented in several forms.

(A) A simple model of the circadian clock gene circuit {Locke, 2005} is shown as an informal diagram, linking four genes (helices) *via* their proteins (ovals), with inputs from light (sun). (B) The differential equation for changes in cytosolic LHY protein (cL_c) in the model is human-readable (and declarative). This equation also involves *LHY* mRNA (cL_m), a translation rate parameter (p_1), RNA degradation rate parameters (m_2, k_2), and translocation of nuclear LHY protein (cL_n) with rates r_1, r_2. (C) A fragment of SBML represents the equation with the same names but is now machine-readable. The first line provides a stable reference to interpret its MathML format. (D) Timeseries simulation of the SBML model in suitable software provided a model output for the RNA level of gene *Y* (Y fit; red, open symbols; timepoints selected to match data), for comparison to RNA data acquired for a candidate gene in Arabidopsis (GI data, filled symbols). After a dark night (-12h to 0h), dawn light transiently induces both the hypothetical *Y* and candidate gene *GI;* the simulation continues in constant light. The comparison of model to data leads to future model refinement (dashed arrow) in the iterative cycle of systems biology. Adapted from {Locke, 2005}.

Figure 2

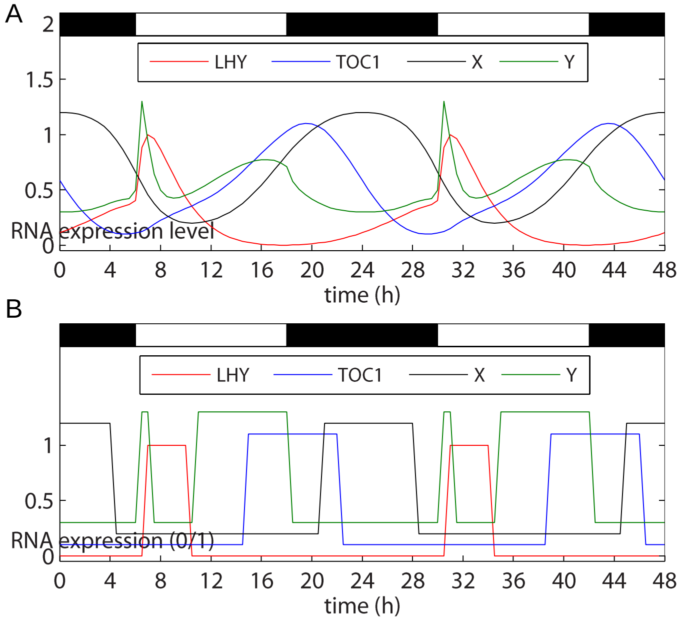


Fig. 2. The simple, qualitative form of a model can retain key behaviours.

(A) Simulation outputs show RNA levels changing continuously, from the simple clock model {Locke, 2005} in quantitative form (differential equations, as in Figure 1B). (B) RNAs are either expressed (1) or not (0) in the qualitative form of the same model {akman, 2012}. The binary, time-delay model still shows bimodal peaks of RNA expression from gene *Y* (green), with light induction after dawn (as in Figs. 1D, 2A). Levels are slightly offset for clarity in (B). Time 0h is midnight. Open box, light interval; filled box, dark interval.

Figure 3

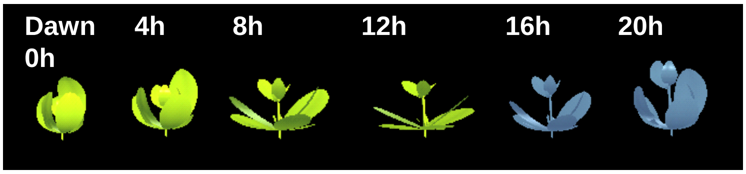


Fig. 3. New capabilities arise from a “black-box” combination of models.

The circadian clock model shown in Figure 1 {Locke, 2005} can communicate to the Arabidopsis architectural model {Mundermann, 2005} running in L-studio software. A version of the clock model in Matlab software was compiled into the C programming language, in order to interact with the lpfg programme of L-studio. A clock protein level from the clock model controlled leaf angle in the architectural model, creating a simple simulation of rhythmic leaf movement in Arabidopsis over day/night cycle. The clock model’s light:dark setting also darkened plant colour at night (16h, 20h). Simulation by Paul E. Brown.

Figure 4

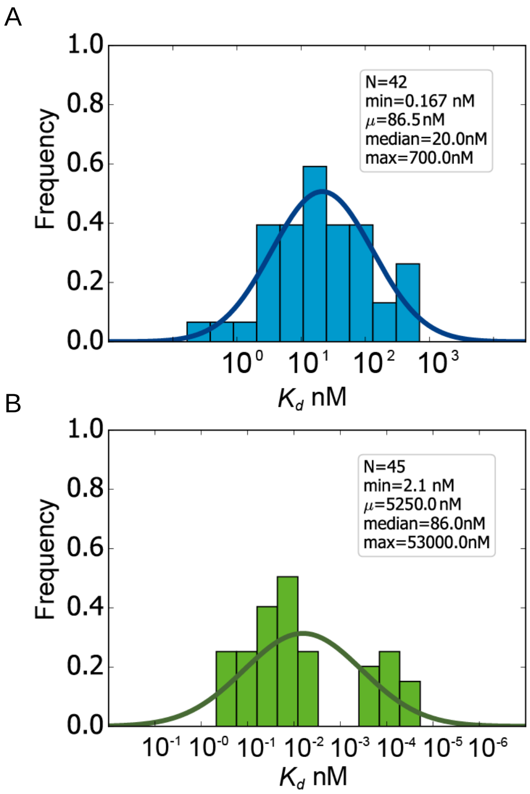


Fig. 4. Published parameter values can inform detailed modelling.

(A) Distribution of published K_d values for plant DNA-interaction affinities. (B) Distribution of published K_d values for plant protein-protein interaction affinities. Data such as these help to constrain the range of parameter values that parameter fitting procedures should explore. Please see main text for publication references.

Figure 5

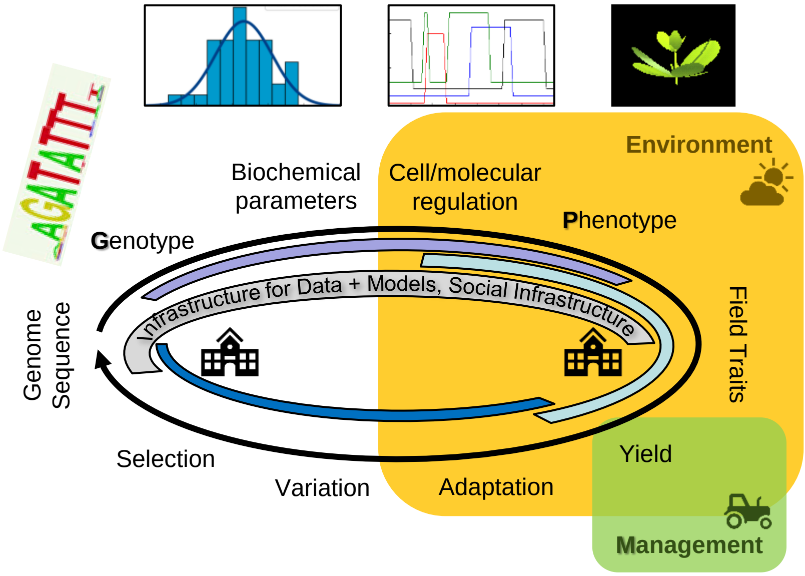


Fig. 5. Linking Systems Biology with Crop Science models.

The solid line links the concepts of biology, first from genome sequence *via* genotype, biochemical parameters and molecular regulation to whole-organism phenotype in a particular environment (yellow area); then from phenotypes to field traits and adaptation or to yield under particular management (green area); finally, given genetic variation, through natural selection or artificial selection in crop breeding, to the evolution of genome sequences {adapted from \Millar, 2016}. Initiatives like Crops *in silico* will deal with the whole cycle, by linking several models (coloured arcs) into a seamless, causal chain. The top line of graphics locate the topics considered in the main text with reference to this cycle. The arcs suggest current types of model, in systems biology (indigo), crop science (cyan) and evolution (dark blue). The dimensions that are often considered in such models are capitalized (G, P, E, M). Underpinning infrastructures (grey) help to bridge these disciplines. ‘Anchor’ institutions are shown (buildings), which might provide major experimental facilities, digital infrastructure or a focus for social infrastructure, such as training or standardisation workshops.