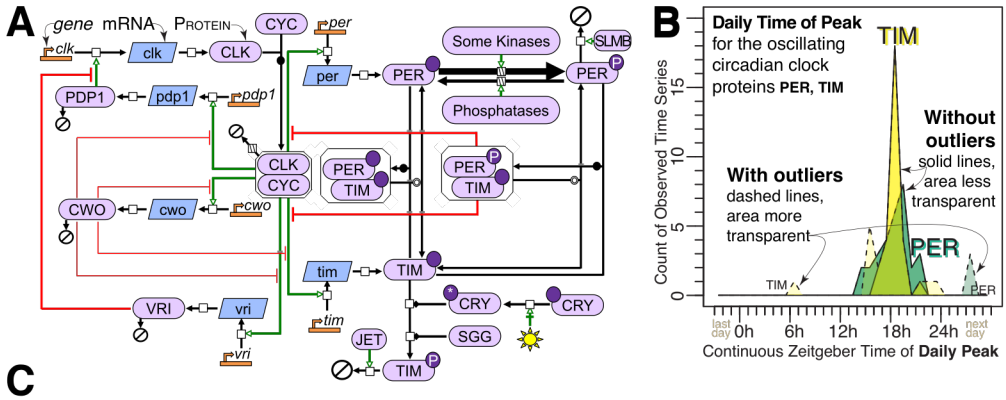
One page overview of core results from the Series '*FlyClockbase and the Future of Biodata Science*'

FlyClockbase, Genome Projects, and other VBIRs

Overview: The FlyClockbase Series consists of interconnected Reports on biodata science that ask key questions about how to integrate diverse research on a given biological system. Using specific examples, each Report provides a different angle on biodata science observations and analyses. FlyClockbase contains >400 circadian clock gene expression time series from the biosystem curation of 25 years of research on clocks in the fruit fly *Drosophila melanogaster* wildtype. FlyClockbase facilitates testing advanced hypotheses about circadian clock components and the methods used to observe them. We use it to show that the clock proteins PERIOD and TIMELESS both peak daily at ~6 PM on average (not new), while the variance for PER is significantly larger than for TIM (our result). We hypothesize that PER's higher phosphorylation complexity adds variability to the time until degradation starts, but other causes may exist. Here we use biodata science to show that our result is not caused by data errors at our end even though we measure our inability to guarantee their absence through human error analysis. More generally, this Series uses **FlyClockbase** to illustrate widely applicable, recurring challenges. This generality inspired our work towards understanding the concepts required for defining generic biodata storage formats that enhance the stability of **Versioned Biological Information Resources** (**VBIRs**). The vantage point achieved by our work on FlyClockbase revealed a panoramic vision of biodata science in which biosystem curation plays the pivotal role of integrating known biological data with a thorough understanding of the underpinning biology, ideally quantifying all relevant uncertainties. We limit our presentation to a panorama of broad terms, which connect key areas for improving analyses of biodata, such as uncertainty quantification, interdisciplinary communication, user-friendly representation of data, compiler construction, statistical logic, and others. Observations we made in FlyClockbase show how a compiler made for biology could improve speed and accuracy of the biodata science required for integrating uncertain knowledge about systems. We developed an interdisciplinary workflow that will enable the construction of such a compiler for handling imperfections of biodata in a user-friendly general-purpose programming language for biology that is made to last.



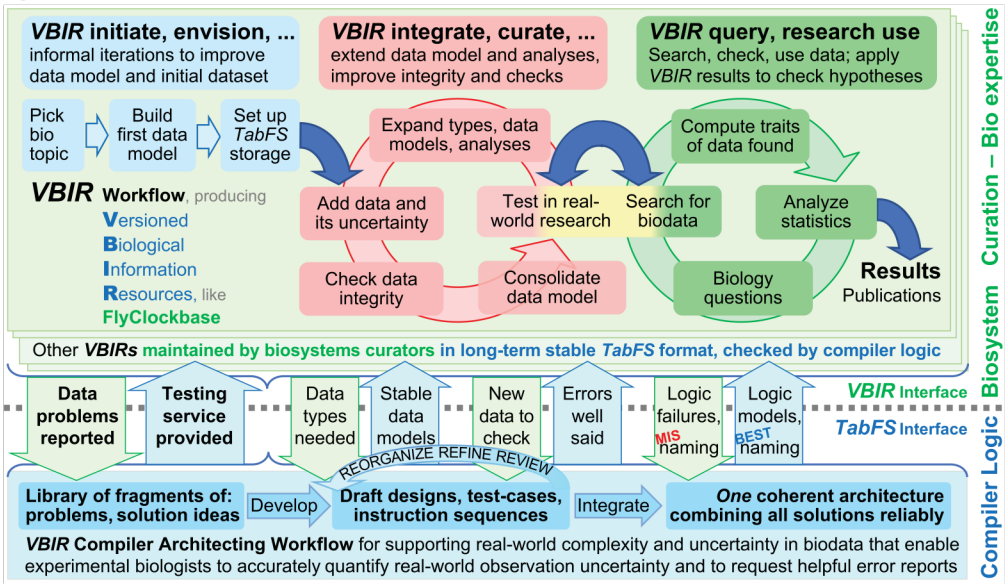


Figure 1: Biodata science integrates various disciplines for improving reproducibility of data analyses. Well-curated VBIRs, like FlyClockbase, boost reproducibility and hypothesis testing speed, like genome projects. We show this by integrating into FlyClockbase 86 studies observing time series of **(A)** wildtype fly circadian clock molecular components, inferring **(B)** peak hours of proteins PER and TIM, revealing different variances. **(C)** Our need for reducing data errors inspired a vision for architecting a compiler that simplifies biosystem curation by integrating and automatically applying expertise from many disciplines for complex error checking as required by imperfect VBIR biodata. To realize this, in-depth bioresearch must meet hard-core compiler design. We illustrate potential forms of such work in biodata science and how appropriate VBIR data structures can help.

VBIR
:
Building Block of Biodata Science

Why VBIRs?
Genome projects convincingly show that batch processing of similar tasks boosts biological research efficiency. Costly reads of single genes shrank to simple queries in the post-genomics era, changing biology profoundly.

Why is **batch-processing** efficient? **It** inspires tools and workflows that speed-up tasks and reuse setup overheads. **It** improves quality by standardization. **It** inspires useful division of labor: a *few* can improve genome quality (via updates), used by *many* for testing hypotheses. Bundling updates into releases helps to improve quality by archiving and citing well-defined genome states reproducibly.

We extend these ideas to other bio data types by introducing the VBIR concept for supporting FAIR data,

Versioned ↔ **F**indable
Biological ↔ **A**ccessible
Information ↔ **I**nteroperable
Resource ↔ **R**eusable,

highlighting rich interactions. Serving its well-defined scope, a VBIR stores all data and updates integrated into reproducibly versioned states of a well-structured biological info resource.

VBIRs vary widely in scope, size, implementation approach, etc. Yet, as indicated by the '**V**', they provide past *versioned variants* via long-term, stable, reproducible URLs. Stable causal VBIRs inspire construction of consequential VBIRs, and help capture complex biological expertise in causality networks. Reproducibility of overall conclusions depends on the stability of VBIR data formats and the reliability of recalculations after auto-importing changed causal VBIRs. Such active networks of VBIRs can infer values, test hypotheses, or simulate complex biological systems. **Compiler-controlled VBIR stability is key for grand research challenges** such as evolutionary systems biology. Compilers can also bundle tasks in batches for improving the efficiency and reproducibility of studies that require biodata science analyses.

For more details, see the initial long argument for biodata science in the FlyClockbase study on BioRxiv at https://doi.org/10.1101/099192

We acknowledge all contributors to work on FlyClockbase as listed there and in this Series.

An earlier variant was presented on 2017-07-04 at the Annual Conference of the Society for Molecular Biology and Evolution (SMBE) in Austin TX. Update v1r7p11 2018-05-21 Available as single page at https://peerj.com/manuscripts/19526/