GP-Write Computing Working Group: Charter and Roadmap

March, 2018

Vision

Data, computing, and identity services to enable synthetic biology at limitless scale.

Mission

The GP-Write Consortium is an ambitious global effort to rapidly advance the state of the art in synthetic biology. Our mandate is to push the boundaries of what is currently possible and to develop the capability to engineer whole multicellular organisms.

In similarly audacious projects (the Human Genome Project, HapMap, ENCODE, TCGA, and others), information technology been both a critical enabler and also a regular source of tension, frustration, and unexpected expense. We expect this to be true for GP-Write as well. While the specific technical challenges of GP-Write will be different from those faced by HGP, we expect many patterns to repeat themselves. *The core lesson of research and scientific computing from the past 20 years is that there must be cross pollination between the science and technology teams.* When we set out to change the world, we must accept that we are changing the rules and requirements as we go along.

This document articulates a set of guiding principles for computing and data infrastructure within GP-Write. It identifies key areas of focus where we believe that technology stands to have a strong impact on the project.

Principles

- *Strong privacy and security practices:* Hold ourselves to the highest standards in terms of data security, privacy, and the protection of personal or sensitive information.

- *Light touch:* Create infrastructure only as needed to advance the consortium's goals. Never impose infrastructure or standards for their own sake.

- *Open source / Open access:* Choose tools, standards, and data formats that will be globally accessible, and for which a free version is available (even if we choose to use the paid version for either support or additional features). This will maximize the community who are able to access our tools and data. We are open to the use of proprietary tools, so long as they do not unduly constrain the usability of the data.

- *Build only as necessary:* Preferentially use existing solutions rather than building new. Avoid "not invented here" syndrome and clearly distinguish between need versus want.

- *Embrace Technology Partnerships:* Offer the best technologies and engineers from business, engineering, media and entertainment, and other industries an opportunity to participate in GP-Write.

- *Avoid technology or vendor lock-in:* Particularly with regard to data storage, ensure that the consortium remains nimble, so that it can take advantage of future innovation and opportunity.

- *Integrate technology and science teams:* The data infrastructure for research cannot be developed effectively across cultural and organizational silos. The teams must share common goals and accountability in order to bring the necessary tempo of innovation.

- *Take responsibility for culture:* Workplace environment, bias, and even conforming to the law are challenges for too many of the leading technology companies. We intend to take social and cultural issues into account when choosing solutions and partners.

Areas of Focus

- *Identity and access management:* Any global collaboration will require significant thought on how researchers, stakeholders, and research participants identify themselves to the project. We expect to use a system of **federated identity**, and to support a **layered security model** so that we apply the appropriate level of protection to each portion of the project.

- *Blockchain:* Distributed ledger technologies allow organizations that do not have a central source of authority or trust to share critical information rapidly and to make decisions in a decentralized way. We see **smart contracts** and **distributed rights management** as areas with the potential for innovative disruption.

- *Machine Learning and AI:* Recent innovations in **deep learning**, and **natural language processing**, coupled with enabling technologies from companies like **NVIDIA** are transforming data intensive fields across multiple industries.

- *Privacy and Information Security:* We expect to address information security in terms of both **policy and practice**. From a policy perspective, we will define strong standards around the **regulatory frameworks** to be followed. In terms of practice, we will take a **modular, layered approach to security,** building protections in at an architectural rather than at some imaginary system perimeter.

- *Permission and consent management:* The effort will certainly include personally identifiable information on research participants. In order to create globally useful datasets, we must **respect local laws and standards** around data privacy, while at the same time integrating ever larger sets of information. We expect that this will require a

system of **machine readable consent**, and also the ability to **re-contact and update** participants while still respecting privacy and anonymity.

- *Data storage and architecture:* In order to be useful, data must be both accessible and also well organized. We anticipate a significant effort to organize and present the project's **metadata** in a comprehensive and scientifically correct framework. This will involve **data modeling** by domain experts including agreement on **ontologies** to allow data to be compared and integrated between researcher groups.

- *Workflows, automation, and reproducibility*: A large part of the consortium's focus is on scaling academic or research pipelines to industrial capacities. This will require automation and process engineering. As part of this, we intend to leverage modern practices around **containerized workflows (Docker and Singularity)** and automated, **data driven analysis (Lambda architectures)**.

Group Membership and Roles

Jason Bobe	Personal Genomes
Brian Bot	Sage Bionetworks
Jack Collins	Frederick National Laboratory
Chris Dwan*	Bridgeplate
Nancy Kelley	NJK Associates
Amy Schwartz	NJK Associates
Nam Pho	NYU Langone Health
Bruce Wilson	Oak Ridge National Laboratory

(*) Chairperson