# Pantograph Research Project

Business Plan

Josiah Seaman - Chief Scientist

## Executive Summary

### Pantograph Product

Pantograph as a project started in 2018 as a way to unlock the next level of population genetics for researchers. Pantograph is a visual browser for Graph Genomes, a new way of capturing sequence data. This process is designed to fix problems common to the sequencing technology that we’ve been using for the past 30 years. For example, Reference Bias means genetic analysis is more accurate for Europeans than for Africans ([Liverpool 2019](https://www.newscientist.com/article/2221957-genetic-studies-have-missed-important-gene-variants-in-african-people/)). A technical paper on Pantograph was being written by a team of 10 scientists before COVID-19 hit.

Now, Pantograph is even more relevant to the current pandemic. The success or failure of our efforts to fight the COVID-19 disease rely upon the sequence diversity of the virus itself. Tests for infection rely on knowing the exact sequence being tested. A rearrangement in the order of genes, even if the content is the same, will return a false negative test.

Second, the vaccine targeting the Spike protein on the outside of the virus relies on a lack of genetic diversity in the Spike protein sequence ([NIH clinical trial](https://www.nih.gov/news-events/news-releases/nih-clinical-trial-investigational-vaccine-covid-19-begins)). If there are any strains with a mutant protein, the vaccine could be rendered ineffective and the virus would continue to spread. SARS-CoV-2 is an RNA virus that will likely infect billions of people, giving it a much higher mutation capability than we’ve previously dealt with in pandemics. For example, the common cold is so impossible to irradicate precisely because of the number of people infected and thus the high number of mutations which exist around the globe. Current sequencing techniques may be under-representing the full sequence diversity of the virus because they are reference based. Eliminating reference bias and enabling species genetic diversity on thousands of individuals is the core goal of using a graph genome.

Pantograph is a very small piece in a worldwide effort to eradicate this disease. It is by no means the most important and the disease will likely end without any involvement on our part. However, given the scale of the pandemic, even a tiny improvement or speedup can result in thousands of lives saved. That’s a difference which is worth investing our resources in. Pantograph will continue to be useful in a wide range of disease application after the current crisis is averted, so that we are never caught unprepared again.

Our short-term goal is to extend the Pantograph tool to add features to make it easier to study the SARS-CoV-2 mutants as the situation develops.

### What is the Pantograph Browser?

Scientists frequently use Multiple Sequence Alignment (MSA) to compare many genomes simultaneously and highlight their differences. MSA works by inserting gaps in all individuals to allow space for unique sequence in one or more individuals. This technique can lead to **Reference Bias** in analyzing diverse sequences. Graph Genomes are a new way of storing an alignment between many sequences that can include **non-linear rearrangements**. For example, the middle of one sequence can be at the beginning of another sequence. Tools for Graph Genomes are still under development and can’t reliably scale to thousands of individuals. Pantograph is the first graph genome browser design with the capability to **scale to thousands of individuals** and still show the individual’s nucleotide sequence. This means it’s uniquely suited to providing a global overview of species genetic diversity with the option to zoom in on small features.

### Future Research

The complete details of Pantograph and its planned capabilities are available in the [Pantograph Specification (Google Document)](https://docs.google.com/document/d/1NEYkRS6Ux1w_v0Soe74FeOAMOxGHOzDun00LdjMi-74/edit?usp=sharing). The key aspects of our approach are:

1. Use Graph Sorting to find a consensus ordering of elements in the genome
2. Separate genome into colinear blocks called Components
3. Show SNPs and indels as cells within a Component grid (Matrix)
4. use colorful Links anchored at Link Columns to show non-linear rearrangements in the pangenome, possibly shared by many individuals
5. Enable zooming from nucleotides, to gene regions, to whole chromosomes by binning of sequence content.
6. Cluster related individuals by sorting the rows into haplotypes

## Organization Description

### Legal Structure

The Pantograph Research Project is under the leadership of the Max Plank Institute research organization.

### Mission Statement

1. Build a tool to track the genetic diversity within a species.
2. Use the tool to understand the variants of COVID-19
3. Create a means for annotating the COVID-19 virus and tracking the patient demographics related to the samples.

### Principal Members

Josiah Seaman

**Institution**: Kew Royal Botanic Gardens, 1001 Genomes Project

**Stage**: Writing up PhD

**Skills**: 12 years Software Development, Data Visualization, Python, Django, C\*, React, Fullstack etc.

**Time Frame**: 3 year grant to build a graph genome browser

**Mainly Interested in:** Visualizing long range links and structural variation, Transposon View

**Projects**: [FluentDNA](http://fluentdna.com/), [DNASkittle](https://dnaskittle.com/try)

Simon Heumos

**Institution**: Quantitative Biology Center (QBiC) Tübingen, Pantograph

**Stage**: Starting PhD

**Skills**: Data Analysis, NGS, Software Development, R, Java, Python, (C++), the only Fullstack experience I had was the knowledge I gained during my master thesis.

**Time Frame**: 2 year grant to build a graph genome browser + 1 year extra = **3 years**

**Mainly Interested in:** Build visualization views supporting clinicians (Trios, Disease Hotspots, Disease Specific Genes, etc.), Read Visualization → Comparison of Mappers using Genome Graphs as a data structure, Extensive Annotation

**Projects**: [TOPAS](https://github.com/subwaystation/TOPAS), [bcellmagic](https://github.com/nf-core/bcellmagic)

Thesis Code, hosted by Computomics:

<https://gitlab.codenic.de/computomics/AGV>

<https://gitlab.codenic.de/computomics/ag>

<https://gitlab.codenic.de/computomics/xg>

Jörg Hagmann

**Institution**: Computomics GmbH, Tübingen, Pantograph

**Stage**: PhD, Bioinformatics Scientist

**Skills**: Data Analysis, (Epi-)Genomics, NGS, R, perl, Python, not really hands-on web dev experience

**Time Frame**: 2 year grant to build a graph genome browser

**Mainly Interested in:** User Friendliness (Performance) = modern, comprehensive visualization and efficient algorithms behind the scenes to serve that need, Integration of other data into graph

**Projects**: Wrote the first read mapping tool against a genome graph ([GenomeMapper](https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-9-r98)), non-related: Pipeline for detecting [DNA methylation differences](https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004920)

Erik Garrison

**Institution**: UCSC, vgteam

**Stage**: PhD, Bioinformatics Scientist

**Skills**: Graph Genome creation, Variant analysis, C++, algorithms

Time Frame:

Mainly Interested in: Graph sorting

**Projects**: author of [vg](https://github.com/vgteam/vg) and [odgi](https://github.com/vgteam/odgi) graph genome models. Contributed to [Sequence Tube map](https://github.com/vgteam/sequenceTubeMap)

Toshiyuki Yokoyama

**Institution**: The University of Tokyo

**Stage**: Starting PhD

**Skills**: Genome Analysis, Software Development, JavaScript, React, Python, Ruby, Docker, Web etc.

**Time Frame**: 3 year grant to build a graph genome analytical pipeline including a graph genome browser

**Mainly Interested in:** Visualizing genome assemblies with gene/repeat annotation and read alignments, establishing a good graph genome format, scalable backend

Projects: [MoMI-G](https://github.com/MoMI-G/MoMI-G/)

URL: <https://github.com/MoMI-G/MoMI-G>

<https://readthedocs.org/projects/momi-g/>

Christian Kubica

Developer Responsibilities

* Simon Heumos - Backend, Project Management, General Overview, SPARQL
* Josiah Seaman - Front-End Design, Project Management, React front end development
* Christian Kubica - Generate Graph genomes to use, Backend graph format specification
* Toshiyuki Yokoyama - Server architecture, Backend graph format specification
* Jörg Hagmann -   
  **Institution**: Computomics GmbH, Tübingen, Pantograph

**Stage**: PhD, Bioinformatics Scientist

**Skills**: Data Analysis, (Epi-)Genomics, NGS, R, perl, Python, not really hands-on web dev experience

**Time Frame**: 2 year grant to build a graph genome browser

**Mainly Interested in:** User Friendliness (Performance) = modern, comprehensive visualization and efficient algorithms behind the scenes to serve that need, Integration of other data into graph

**Projects**: Wrote the first read mapping tool against a genome graph ([GenomeMapper](https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-9-r98)), non-related: Pipeline for detecting [DNA methylation differences](https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004920)

Erik Garrison

**Institution**: UCSC, vgteam

**Stage**: PhD, Bioinformatics Scientist

**Skills**: Graph Genome creation, Variant analysis, C++, algorithms

Time Frame:

Mainly Interested in: Graph sorting

**Projects**: author of [vg](https://github.com/vgteam/vg) and [odgi](https://github.com/vgteam/odgi) graph genome models. Contributed to [Sequence Tube map](https://github.com/vgteam/sequenceTubeMap)

Toshiyuki Yokoyama

**Institution**: The University of Tokyo

**Stage**: Starting PhD

**Skills**: Genome Analysis, Software Development, JavaScript, React, Python, Ruby, Docker, Web etc.

**Time Frame**: 3 year grant to build a graph genome analytical pipeline including a graph genome browser

**Mainly Interested in:** Visualizing genome assemblies with gene/repeat annotation and read alignments, establishing a good graph genome format, scalable backend

Projects: [MoMI-G](https://github.com/MoMI-G/MoMI-G/)

URL: <https://github.com/MoMI-G/MoMI-G>

<https://readthedocs.org/projects/momi-g/>

Christian Kubica

## Market Research

### Industry

Describe the existing systems and the need for Pantograph.

A tool for tracking all strains of the virus

* genetic diversity in large population
* graph structures naturally express genome rearrangements
* provide common communication around genetic knowledge, places, and concepts
* viral strains can be smoothly integrated as they are sequenced

### Detailed Description of Customers

Who needs this research?

Current tools do not scale

* tools that are scalable currently discard all genome rearrangement
* must account for variation across population
* universal standard data structure that scales to millions of individuals with all variants
* support integrate of data from many sources

### Company Advantages

Highly effective team of volunteer with the proven track record. This was demonstrated int he Bio-hackathon.

### Regulations

All of our research and software will be available globally under the Apache licensing agreement.

## Project History

### Product/Service

Pantograph has its root in the Fluent DNA project which began in 20xx.

It also builds on the research of Pangenome Browsers

Pantograph has been used to display genetic diversity for researchers since 2020?

### Pricing Structure

All of our research and software will be available globally under the Apache licensing agreement.

### Product Lifecycle

A number of milestones

* Basic release
* Hackathon phase - performance improvements
* Graph Data Structure - to support genetic annotation

### Intellectual Property Rights

Wooden Grain Toys is a trademarked name in the State of Oregon.

### Research and Development

Danger of surprise mutations

* different viruses have different mutation rate and duration of immunity
* mutants can vary in infectiousness and mortality
* infected population size is multiplier to mutations available for selection
* vaccinating the population during a pandemic selects for mutant resistance
* this is the biggest selection sweep of a virus in human history

Vaccine deployment effectiveness requires knowledge of genetic diversity

* Vaccines will not be equally effective against all strains
* Vaccine effectiveness is tied sequence similarity to the vaccine sample
* Pantograph can help predict vaccine effectiveness in different regions of the world

## Marketing & Engagement

### Growth Strategy

1.0 Release - Initial release and promotion

1.1 - First update milestone with performance and quality improvements

1.2 - Enhanced data collection capabilities

2.0 - Semantic data and annotations

### Communicate with the Other Researchers

Researchers are actively engaged on Twitter among the scientific community.

### Future Promotion and Collaboration

What is being done to make these tools available and create awareness?

### Data Collection and Collaboration

We currently have an infrastructure for gathering and uploading sample of the SAR-COV2 virus for clinical sources.

We have defined a data governance model to protect the data integrity.