Part I - Introduction Slide Deck

**Slide 1**: Welcome to the Genome Solver Project.  My name is Anne Rosenwald and I’m a professor of biology at Georgetown University.  This slide deck discusses rationale for the project. Our overall aim is to engage biology faculty in basic bioinformatics, teaching some of the online tools available for doing authentic research in this area with students.  What we hope to convince of with this series of 6 short powerpoint slide decks and associated exercises is that the bioinformatics tools are easy to find and use, the DNA sequence data are freely available, so hopefully you will think about incorporating some aspects of bioinformatics in your classes.

**Slide 2:** The Genome Solver project has been funded by the National Science Foundation for  6 years now.  The project was initially developed by me and then post-doctoral fellow Gaurav Arora, now at Gallaudet University, and continues today with Vinayak Mathur, another former post-doc who is currently at C.

Up until the summer of 2017, we conducted workshops for faculty at different colleges and universities around the country.  Now that the funding is coming to an end, we have made the current workshop materials available so that others can learn, too.

**Slide 3**: The impetus for the Genome Solver Project came initially from the Human Microbiome Project an effort funded by the NIH to understand the ecology of microbes that live in and on us. There is a wealth of DNA sequence available - there are about 45,000 sequenced bacterial genomes available now. If we assume that each genome is the same size as E.coli, which is a rough approximation, there are gigabytes of data available for the students to play with. Much of this has only been annotated by machine - meaning that a computer program has looked at it and predicted genes but a lot can be learned by looking at it more deeply and comparing sequences among different organisms.  In addition, there are collections of 16S rRNA sequences available, describing population structure of bacteria in different environments.  We won’t talk much about this kind of analysis but that information is also available

What we are going to talk about in this series of short slidedecks is how you can do comparative genomics using these freely available data -  your students can look at these data and make inferences about the function of interesting genes and how they vary over the course of evolution. I’ll introduce some published projects done in my lab with undergraduate authors. So we hope to encourage you to use some of these simple tools available on the internet and you too can get publishable work out of it in a way that is good for your CV and your students.

This graphic shows some of the data that has come out of the human microbiome project.  In 2012 there were  two anchor papers that came out in Nature, and a set of 14 other papers that came out in a variety of different journals. The message was that different parts of your body have different microbes associated with them so there are these different ecological niches. A lot of the students come to college thinking about medical school, and the microbiome studies are a way to sneak some ecology into the curriculum. Each body part has a different microbiome so this is a great hook to get students to start thinking about ecology a little bit.

**Slide 4**: Here is our idea: We have this data fire-hose and this fire-hose presents an opportunity. There are more data there than bioinformaticians are really looking at. This allows students to get involved in genuine research projects and they can learn some tools at the intersection of biology, math, computer science and statistics.

They can develop hypotheses based on the genome information. Some of the work we have done uses these hypotheses to test predictions in the wet lab. Overall we believe this is a win for everybody: bioinformaticians, students and faculty because there are data that anyone can look at and we can find out new interesting things about biology, expanding what we know about life processes.

**Slide 5**: Our learning goals for you are: to gain some facility with the tools that we will talk about today. It is important to preface this by saying that the tools that we are going to show you are not the only tools that are available, these are the ones we like and we think they are easy for the students to use. There are many different ways to do different things. We would like you to become familiar with our website: https://genomesolver.qubeshub.org. We hope you can connect your curricular needs with your research needs. In the last video in the series we will talk about our community science project.

**Slide 6:** The other materials in this series are as follows

•Part II:  Databases

•Part III: Gene Annotation

•Part IV:  Comparative Genomics

•Part V:  Phylogenetics

•Part VI:  Community Science Project

For each of the subsequent slidedecks except for part VI, there is an associated exercise to practice the tools introduced.  These are available on genomesolver.qubeshub.org

**Slide 7**: What we are going to do in this series is replicate what we did in a paper published back in 2014. This will introduce the idea of the community science project which we’ll discuss in more detail in Part VI. Brad, Theo, and Alex are former undergraduates who are authors on this paper. In a class that Gaurav and I taught in the Spring of 2012, Theo found an instance of a bacteriophage gene embedded in a bacterial genome. It turns out that there are a lot of phage genes in many bacterial genomes - this forms the basis of the community science project - identifying orthologs of phage genes in bacteria. So what we will do is learn how to find genes, learn how to find the orthologs of those genes in other species or other bacteriophages.  Then we will do an alignment and create phylogenetic trees.

**Slide 8**: Our big scientific question is what are these phage genes doing in bacteria. For many of these bacteria, we can’t grow them in the laboratory and even if we could, there might not be a genetic system available.  Therefore, bioinformatics is our only way to investigate this interesting phenomenon.

To begin, please learn about Databases with the 2nd slidedeck in the series.