Genome Variant Analysis 2017

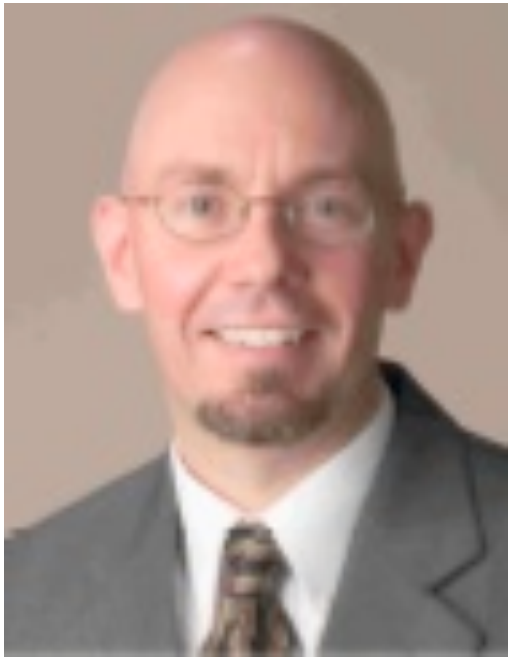
General Introduction

Background

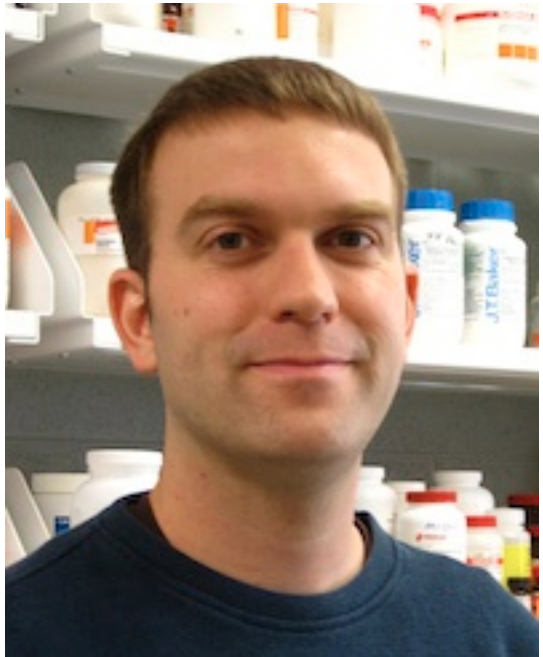
- Key research interest identifying rare variants in variety of systems and determining how to leverage that information.
- Self-taught computational biologist
- TA two years
- 4th year teaching this course

A nod to the past …

- Scott Hunicke-Smith
- Former Director of Genome Sequencing Analysis Facility



- Jeffrey Barrick
- Assistant Professor MBS & ICMB



Disclaimers up front

- Royal “we” – Tutorials written without clean tense and pronouns. We likely means me, so don’t blame Dacia for any problems you find, but feel free to assume anything that is working is only because of her work.
- Spelling – Has never been a skill I possess, hopefully will only be noticeable if I write on board.
- Typos – Will likely be your biggest problem in using the commands we provide, and the biggest problem in your own work. Try to type the commands out to get practice, but remember you can copy paste.
- Names – I usually use this opportunity to apologize for my inability to remember people’s names with a funny anecdote, with small class might not be an issue.

Goals and an early attempt to manage expectations.

- Participant goals:
  - Learn how to analyze my data, and have it fully analyzed by the end of the class.
  - Learn how to analyze NGS data in general.
- Teaching goals:
  - Teach the fundamentals of NGS variant analysis.
  - Provide context and exposure multiple types of data.
  - Use example commands to familiarize you with variety of programs.
  - Provide resources to enable you to do analysis you haven't thought of yet.

Where to start

- Many say “don’t know where to start” their data analysis once they have their data files.
- Ideally should have “started” weeks-months ago in planning experiments.
  - If you fall in this category don’t worry you are in a common situation.
  - Later presentations will explain differences in sequencing libraries and how to make the best choices.

General Impressions of field

1. There is no “standard” or “best” way to analyze next-gen data.
    - Results in many seemingly equivalent or redundant analysis types or tools.
    - Not all types of analysis are appropriate or even necessary.

2. Wet lab methods used to generate data may limit available/appropriate analysis methods.
    - Not-surprisingly, planning experiments from the beginning based on what answers/analysis you want at end can save time and money.

Computers Computers Computers

- Millions of reads, 100s of bp long, mapping to millions-billions of base long references.
- Windows is your enemy, linux/Mac is your friend.
- TACC is a time machine that lets you get stuff done much faster
  - This is where we will start the class.