



Introduction

- [Your Instructors](#)
- [Communication](#)
 - [Asking questions](#)
 - [Getting help](#)
 - [Conventions](#)
- [Course goals](#)
- [NGS Challenges](#)
 - [Diverse skill set requirements](#)
 - [Large and growing datasets](#)

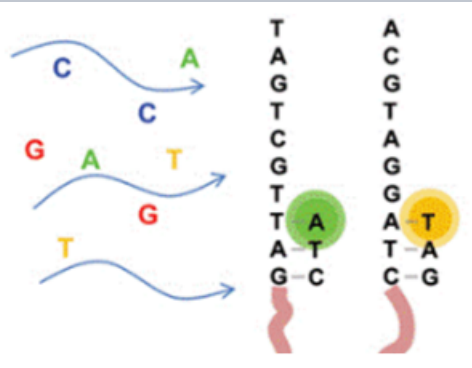
Your Instructors

- **Anna Battenhouse**, Associate Research Scientist, abattenhouse@utexas.edu
 - BA English literature, 1978
 - Commercial software development 1982 – 2007
 - Joined Iyer Lab 2007 ("retirement career")
 - BS Biochemistry, UT Austin, 2013
 - Joined the Biomedical Research Computing Facility ([BRCF](#)) and [Marcotte Lab](#) summer 2017
 - Also affiliated with
 - Bioinformatics Consulting Group ([BCG](#))
 - Genome Sequencing and Analysis Facility ([GSAF](#))
- **Daryl Barth**, daryl.barth@utexas.edu
 - BS Materials Science & Engineering, UC Berkeley, 2017
 - Student Researcher in France and Portugal 2017 - 2018
 - Research Assistant in single cell genomics, UT Southwestern 2019-2021
 - 3rd year graduate student in the [Marcotte Lab](#)
 - Research Interests: biomaterials, developmental biology, and bioinformatics

About the Iyer Lab (where Anna learned NGS)

http://iyerlab.org/ Dr. Vishy Iyer, PI	
Main focus is functional genomics <ul style="list-style-type: none">◦ large-scale transcriptional reprogramming in response to diverse stimuli◦ Encode consortium collaborator◦ works in human and yeast	
Research methods include <ul style="list-style-type: none">• microarrays (Dr. Iyer was co-inventor)	

- high-throughput sequencing (since 2007)
 - especially ChIP-seq, RNA-seq
 - also miRNA-seq, RIP-seq, MNase-seq ...
 - have ~2,000 NGS datasets



Communication

Asking questions

Feel free to ask questions any time during the instructor's lecture and demonstrations.

For online attendees, you can also post your question to the Zoom chat. We'll sometimes use breakout rooms when troubleshooting problems you run into, if so, TA Daryl Barth will assign you to one.

Getting help

Since most folks are new to the Linux command line, we expect you to run into problems! Please let us know if you're having difficulties!

Making mistakes and running into problems is key to learning the Linux command line! It is not only expected – it is encouraged 😊.

Conventions

If you see a block of text like this:

Example code block

```
ls -h
```

it means, ***type the command `ls -h` into a terminal window, hit **Enter**, and see what happens.***

We intend this course to offer as much self-learning as possible. Consequently, you'll find many sections like this - click on the triangle to expand them:

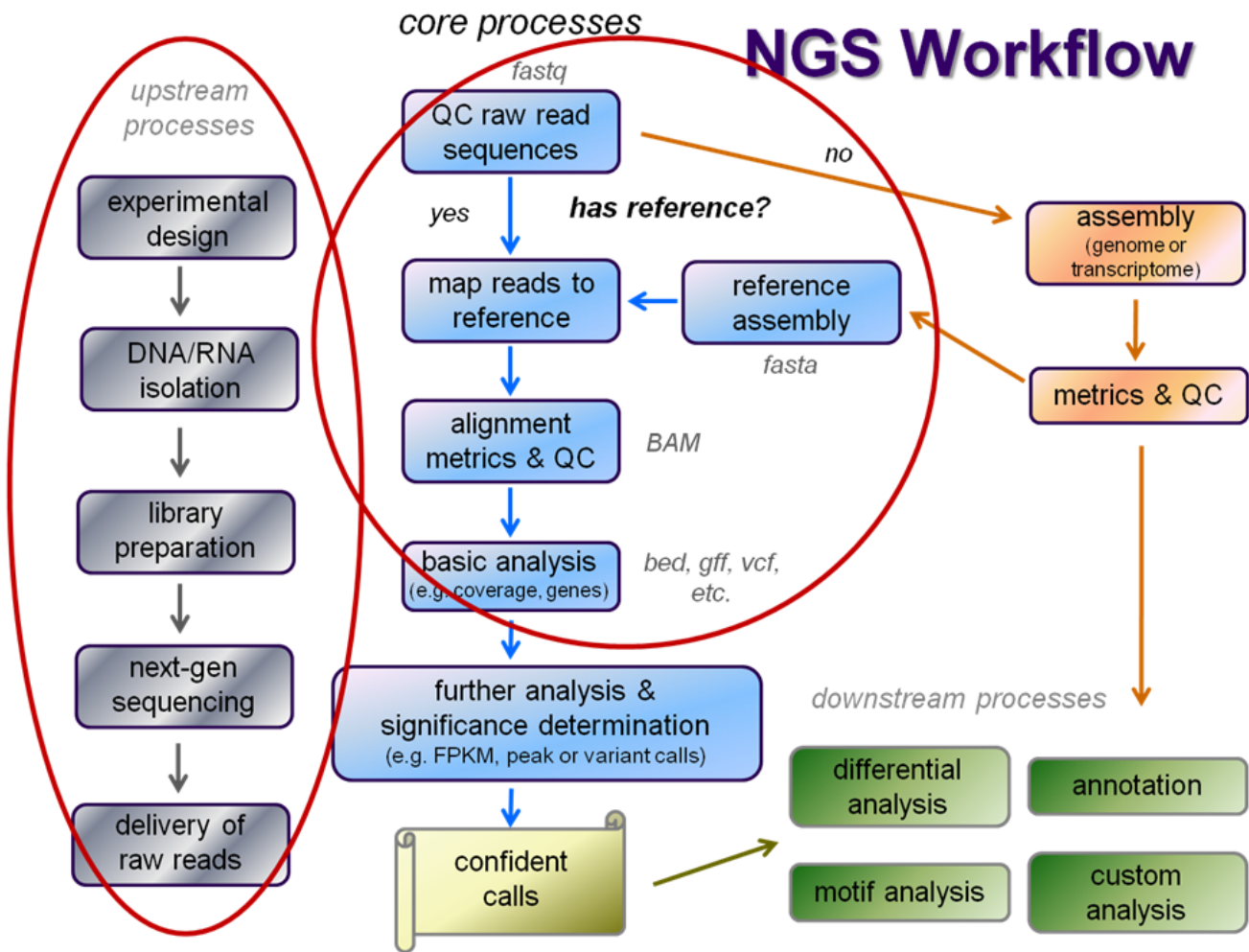
Hint sections will provide you some guidance on what to do next, but will not spell it out.

and some sections like this:

Solution sections will contain the `commands` so that you could copy-and-paste them if you have to. They will represent one method of answering the question – but there are often many ways to skin a cat!

Course goals

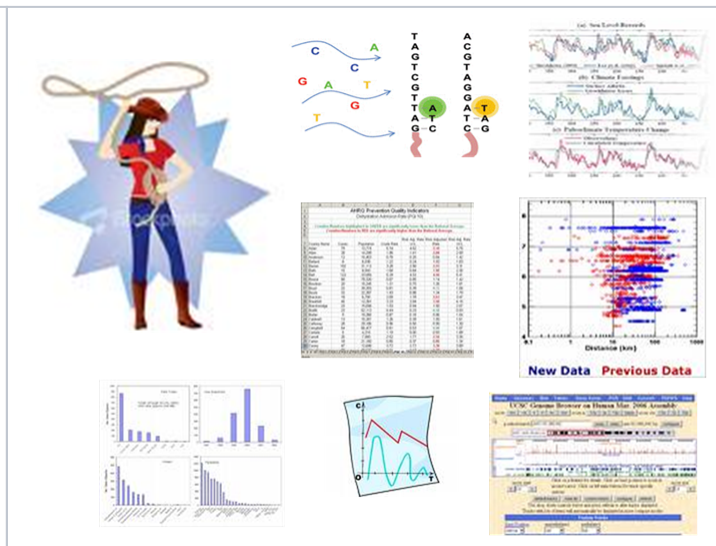
- Hands-on, tutorial style – learn by doing
 - Common bioinformatics tools & file formats
- Introduce NGS vocabulary
 - both high-level view and practice with specific tools
- Cover the NGS basics
 - The first few things you'll do after receiving raw sequences
 - raw sequence QC and preparation
 - alignment to reference
 - basic alignment analysis
- Understand and practice required skills
 - Get you comfortable with Linux and TACC – your best "frenemies"
 - Make you self-sufficient enough in 5 days to become experts over time
 - Show some "best practices" for working with NGS data



NGS Challenges

Diverse skill set requirements

- **Analysis** – making sense of raw data
 - one part bioinformatics and statistics
 - one part scripting / programming
 - Linux command line
 - High Performance Computing (TACC)
 - **bash** scripting (**grep**, **awk**)
 - **R**, **python**, **perl**
- **Management** – making order out of chaos
 - one part organization
 - one part data wrangling
- **Adoption of best practices is critical!**



Large and growing datasets

NGS methods produce staggering amounts of data!

Typical dataset these days

- yeast: 5 – 20 million reads
- human: 20 – 250 million reads (~5 - 8 million for TagSeq)
- single end (SE) or paired end (PE), length 50 – 300 bases (100 or 150 typical)

The initial **FASTQ** files are big (100s of MB to GB) – and they're just the start.

- Organization and naming conventions are critical.
- Your data can get out of hand very quickly!

Progression of Iyer Lab datasets over time:

- 2008 – Yeast heat shock remodeling of chromatin
 - 2 yeast datasets
 - less than 2 million sequences
- 2010 – Allelic bias in CTCF binding
 - 13 CTCF datasets from 3 GM cell lines
 - ~200 million sequences
- 2012 – Transcription factor data analysis (ENCODE2)
 - 32 ChIP-seq datasets gathered over 3 years (3 TFs across 11 cell lines)
 - ~ 1 billion sequences
- 2013 – miRNA overexpression effects
 - 42 RNAseq datasets (7 conditions)
 - ~ 2.6 billion sequences
- 2014 – eQTL analysis of CTCF binding
 - 52 very deeply sequenced CTCF datasets
 - ~ 8 billion sequences
- 2018 – Functional analysis of glioblastoma tumors and cell lines
 - nearly 500 datasets in total (ChIP-seq, RNAseq, miRNAseq, 4C, exome/genome sequencing)
 - > 22 billion sequences