NCGR Advanced Bioinformatics

Prepared for
University of Alaska, Fairbanks and Anchorage
January 28-29, 2016

Faye D. Schilkey
Director of Strategic Projects and
NM-INBRE Sequencing and Bioinformatics Core
Agenda

- NCGR
- Education
- Research
- Sequencing Center
- Bioinformatics
- Understanding Projects at UA F & A
National Center for Genome Resources
501c(3) nonprofit research institute

Applies bioinformatics, software engineering and next-generation sequencing to solve the preeminent challenges of 21st century biology
NCGR History

- 1994 NCGR founded for discovery-driven computational biology and bioinformatics in support of the DOE Human Genome Project
  - Created 1st *relational* human genome sequence database (GSDB)
- 1997-98 NCGR established and sold Molecular Informatics, now part of Celera Diagnostics
- 2000 NCGR’s completes new 32,000 square foot research facility
2000-2007 NCGR continues innovative bioinformatics tool development

- SYStem for Integrating heterogeneous bioinformatic resources (ISYS)
- Comparative Map and Trait Viewer (CMTV)
- A metabolic database and discovery tool (PathDB)
- A tool for Genomic Exploration and Survey of Immune Response (GEYSIR)
- The *Arabidopsis* Information Resource (TAIR)
- The Legume Information System (LIS)
NCGR History (cont.)

- 2007 NCGR creates NM Genome Sequencing Center
  - Alpheus created to analyze Next Gen Sequencing (NGS) data
- 2007 – date, Major sequencing & analysis projects
  - NIH NM-INBRE
  - Moore Foundation Marine Microbial Transcriptome Sequencing Project (MMETSP)
  - NSF Medicago HapMap project
  - Cacao Genome project funded by Mars, Inc.
  - Development of a human carrier screening test
  - NOAA – Deep Well Horizon
- 2014 NCGR develops Bioinformatics-in-a-Box™
Faye D. Schilkey, BS Computer Engineering

- First Career: Software engineering in automotive (robotics) and aerospace (guidance and autopilot) systems

- Second Career (Big Data):
  - IT / Software Engineering / Database development in Genomics and Bioinformatics (20 yrs)
    - GSDB, ISYS, PathDB, GEYSIR
  - PI/Director, NMINBRE Seq & Bioinformatics Core (10 yrs)
  - Sequence Center operations, R&D, and product development (9 yrs)
  - Leverage sequencing and bioinformatics techniques to solve biological questions
Main elements of NCGR

- Bioinformatics
- Sequencing
- IT and Software engineering
- Research
- Education
  - STEM / undergrad / grad
2004-2015 period > 2500 researchers/students positively impacted through SBC activities

- Sequencing/bioinformatics for projects, pubs & grants (215)
- College genomics/bioinformatics internships (125)
- Pre K - grade 12 genomics/bioinformatics education (722)
- BioInformatics, Science and Technology Symposium-NMBIST (1080)
- College genomics/bioinformatics seminars (525)
Research at NCGR

- Human health
- Plant science and nutrition
- > 200 publications
Human Health Research

- Dengue virus infection (Virology 2015)
- Vibrio cholerae (Genomics Discovery 2014)
- Guinea Pig (Genome Announc 2013)
- Eyeless Hedgehog (PLoS One 2012)
- Carrier Screening (Beyond Batten - Sci Transl Med 2011 )
- Multiple Sclerosis (Twins study - Nature April 29, 2010 cover)
- Sepsis (J Clin Microbiol. 2010)
- Korean Genome (Nature 2009)
- Mesothelioma (Proc Natl Acad Sci 2008)
• **Medicago truncatula** (Barrel clover) **HapMap** (500Mb)
  - Cornell, UVM, JCVI, NSF, UCSC, INRA-Montpellier, ENSAT-Toulouse, Boyce Thompson Inst.
  - Samuel Roberts Noble Foundation

• **Medicago sativa** (Alfalfa) Genome (860Mb)
  - Samuel Roberts Noble Foundation

• **Theobroma Cacao** (Chocolate) Genome (330Mb)
  - USDA-ARS & Mars, Inc., Washington State University, JGI, USDA-ARS, IBM, PIPRA, CUGI

• **Glycine Max** (Soybean) (1 Gbp) and **Zea Mays** (Maize) (2Gb) **Genetic Diversity**
  - Syngenta

• **Sorghum Transcriptome**
  - USDA-ARS

• **Gossypium arboreum** (Cotton) Genome (1.7 Gbps)
  - Texas Tech University & Bayer Crop Sciences

• **Phytophthora capsici** (Chile pepper blight) (100 Mbps)
  - Univ. of Tennessee, Ohio State Univ., USDA/NSF

• **Legume Disease Resistance**
  - National Science Foundation, University of California – Davis

---

**Plant/Animal/Fungi/Bacteria Science**
• **Chickpea & Pigeon Pea Diversity**  
  – CIMMYT - Generation Challenge Program, ICRISAT

• **Andean Birds (Hummingbird) Transcriptome** (1 Gbp)  
  – UNM, NSF

• **Green Microalga** (85 Mbp) and **Diatom strain RGd-1** (25 Mbp) **Genomes**  
  – Center for Biofilm Engineering, Montana State University

• **Staphylococcus aureus strains** (3 Mbp)  
  – NMSU, OSU, NIH, NM-INBRE

• **Burkholderia glumae (rice blight) genome** (7.3 Mbp)  
  – Louisiana State University

• **Bacteroides xylanisolvens strains** (6 Mbp)  
  – USDA-ARS, DARPA, Vital Probes

• **Polaromonas sp. Strain CG9_12 (pollutant degradation) Genome** (5 Mbp)  
  – Center for Biofilm Engineering, Montana State University

• **Kibdelosporangium sp. MJ126-NF4** (Actinobacteria having natural products: anti-bacteria/viral/cancer) **Genome** (11 Mbps)  
  – UNM

---

Plant/Animal/Fungi/Bacteria Science (cont.)
NM INBRE Collaborations

2015 INBRE

2008-2014 INBRE & HHMI SEA

Q1 2016: 6 planned
CS-Pro certified Sequencing Center

Cutting-edge collaborative research and service projects featuring DNA and RNA sequencing spanning all kingdoms of life

- Genomes of any size (re-sequencing or de novo assembly)
- Genomic target selection (e.g. exomes or amplicons)
- ChIP-Seq
- RNA-Seq (PolyA selection or RiboZero)
- Small RNA sequencing
- Genotyping-By-Sequencing (GBS)
- 16S/18S or whole genome metagenomics
- DNA methylation

• PacBio IsoSeq full length RNA sequencing
Broad Bioinformatics Arsenal

- Microbial community analysis
- Assembly of simple and complex genomes
- Transcriptome assembly and annotation
- Differential gene expression
- Variant discovery
- Structural Variation
- Pangenomics
- Epigenetics and ChIP-Seq
- Small RNA analysis
- Emerging technologies

Areas of interest?
Microbial community analysis

Approach: Amplification, sequencing and analysis of the variable regions of the 16S (ribosomal small subunit) segment of bacterial and archaeal ribosomal RNA cistrons

- NCGR has experience in both bacteria and fungi communities using ribosomal genes
- Currently working on whole metagenome PacBio sequencing and analysis of low complexity microbial communities
Assembly of simple and complex genomes

Pros: *De novo* assembly of sequence read data into representations of the molecules sampled avoids biases of interpretation inherent in read mapping to existing references.

Challenges: Can be complicated due to highly repetitive elements, low complexity regions, coverage biases, recent duplications and allelic heterozygosity.

- **NCGR has experience:**
  - From simple bacteriophage through to polyploid higher plants, applying short and long read technologies independently and in combination.
  - Improving genome assemblies through incorporation of genetic and physical maps, including the newly available BioNano Irys system. E.g. cotton genome work with Dr. Joshua Udall.
BioNano Irys Technology

http://www.bionanogenomics.com/technology/irys-technology/

Background: Images HMW motif-labeled DNA and uses algorithms to convert into molecule maps. These maps provide dense genome-wide anchor points for ordering and orienting contigs or scaffolds to improve NGS assemblies.

- Retaining long-range contiguity throughout the genome mapping process provides a comprehensive study of genome structure and function.
- In particular, de novo sequence scaffolding and analysis of structural variation in complex genomes.
- Structural variants and repeats are measured directly within long, single-molecule “reads” for comprehensive analysis of what has been dubbed “the inaccessible genome.”
Transcriptome assembly, annotation and landscape

Background: Greatly smaller in size relative to the genome.

Challenges: 1) Large range of relative abundances of the molecules sampled
2) Presence of identical sequence segments in differentially spliced or post-transcriptionally modified forms of transcripts

- NCGR has a fully tested, high throughput pipeline for assembling mRNA reads into their cognate transcripts and annotating these with respect to protein motif databases
- NCGR offers PacBio IsoSeq which provides reads that span entire transcript isoforms, from the 5' end to the 3' polyA-tail to directly sequence full-length transcripts ranging up to 10 kb. No assembly required!
Differential gene expression

Background: Determining differences in gene expression profiles among mRNA samples using Illumina sequencing has essentially replaced fixed probe array-based methods

- NCGR has extensive experience in applying methods to discover statistically significant differentially expressed genes in a variety of contexts including:
  - allele-specific expression
  - differential splice form quantification
  - pathway and gene ontology term over representation
  - neo and subfunctionalization of duplicated gene families
  - integration with other quantitative omics measurements such as ChIP-Seq of transcription factor binding and methylseq data
Variant discovery

Background: Single nucleotide polymorphisms (SNPs) may have no effect on the amino acid they produce while others are deleterious.

- NCGR was a pioneer in variant discovery in next-generation sequencing.
  - Awarded the 2009 BioIT World Best Practices Award for research into the genomics of schizophrenia, in which it employed its Alpheus® variant detection pipeline.
  - Alpheus also was recognized by the 2009 Computerworld Honors Program.
Selection of patient cohorts

- Sequencing
- Discovery of variants
- Association of variants with disease traits in patients
Structural Variation

Background: Structural variants can have major impact on phenotypes but are more difficult to ascertain using reads that are short relative to the event lengths.

- NCGR has experience in development and utilization of specialized methods for detection of specific subtypes of structural variation in targeted regions and genome-wide using:
  - pairing signatures
  - mapping technologies
  - assembly-based methods

- NCGR is exploring the use of physical mapping strategies (e.g. BioNano) for identifying large structural variations.
Pangenomics

**Background:** Pangenomics attempts to describe the diversity of and relationships among elements of genomic content within a species, breaking away from the established paradigm of representing the entire species by describing a single individual.

- NCGR scientists have several ongoing efforts to explore methods for efficient and meaningful characterization of species with multiple sequenced genomes.
Epigenetics and ChIP-Seq

Background: Epigenetic changes are critical in mediating organismal responses to environmental changes and in modulating the differentiation of cell types during development.

- NCGR can provide 5-methylcytosine analysis of bisulphite converted genomic DNA sequenced on Illumina as well as ChIPSeq assessment of regions of DNA-protein interaction.

- Additionally, the PacBio RS instrument can detect several base modifications based on the kinetics of nucleotide incorporation.
Small RNA analysis

Background: Noncoding RNAs are an important part of the transcriptomic landscape, exerting an important layer of gene control.

- NCGR has experience in quantitative and differential profiling of miRNAs as well as target prediction.
Emerging technologies

Background: NCGR is particularly interested in developing expertise in activities at the cutting edge of genomics and bioinformatics.

- Areas of exploration include:
  - Single cell plant omics
  - New and emerging physical/optical mapping strategies
  - NCGR is working with DoveTail Genomics to beta test their Chicago libraries in crop plants
  - Nanopore MinION
Acknowledgments

NCGR/NMINBRE Sequencing and Bioinformatics Core

Science/Bioinformatics
Anitha Sundararajan
Johnny Sena
Joann Mudge
Nico Devitt
Thiru Ramaraj
Stephanie Guida
Connor Cameron
Andrew Farmer
Diego Fajardo
Boris Umylny
Callum Bell

Sequencing Lab
Peter Nagm
Jennifer Jacobi
Pooja Umale

IT/Administration
Forrest Black
Kathy Myers
Lisana Chavez

NIH
NIGMS (5P20GM103451)
Announcements

New Mexico BioInformatics, Science and Technology (NMBIST) Symposium on

“Advances in Genome Technology”
March 17,18 2016
Santa Fe, NM

- Experts in the field
- Student poster session
- Student speaking slot competition

inbre@ncgr.org
Thank you!

Please reach me to scope your NGS and/or analysis project!

Faye D. Schilkey
505-995-4449
inbre@ncgr.org