Course Overview:
Mutation Detection Using Massively Parallel Sequencing
From Data Generation to Variant Annotation

Outline

• Course Overview:
  • Who we are: Introduction of Course Organizers and Lecturers
  • Who you are: Background of Participants
  • Day-by-day
  • Tools for the course

• Introduction to Massively Parallel Sequencing
  • Whole Genome, Exome Sequencing, and Targeted Sequencing
  • Targeted Genomic Enrichment and Multiplexing
  • Examples
Richard Smith, MD
Director, Iowa Initiative in Human Genetics
Sterba Hearing Research Professor
Professor of Otolaryngology, Internal Medicine, Pediatrics and Molecular Physiology and Biophysics

Tom Bair, PhD
Director, Bioinformatics
Iowa Initiative in Human Genetics
Associate Research Scientist, DNA Facility

Ann Black-Ziegelbein
Senior Application Developer
Center for Bioinformatics and Computational Biology and Iowa Initiative in Human Genetics

Eliot Shearer
Graduate Student
Medical Scientist Training Program (MD/PhD)
Department of Molecular Physiology and Biophysics

Ben Darbro, MD PhD
Assistant Professor in Pediatrics & Laboratory Director of Cytogenetics and Molecular Pathology, UIHC

Kevin Knudtson, PhD
Director, DNA Facility Research Division, Iowa Initiative in Human Genetics

Ben Rogers, PhD
Director, Research Services University of Iowa ITS

André Altmann, PhD
Postdoctoral Scholar Stanford University

David Dimmock, MD
Assistant Professor in Pediatrics Human and Molecular Genetics Center, Medical College of Wisconsin

Alex Nord, PhD
Postdoctoral Scholar Lawrence Berkeley National Laboratory, Joint Genome Institute
Course Participants

- Participant Classification
  - Masters/Medical/Graduate Student: 18%
  - Postdoctoral Fellow: 12%
  - Professor: 14%
  - Research Assistant/Manager: 14%
  - Research Scientist: 34%

- Broad Range of Experience

Experience with Galaxy?

- Yes: 62%
- No: 38%

- Majority have not used Galaxy

Overall Goals for the Course

- To provide a broad overview of massively parallel sequencing (MPS) technology
- To provide a framework for investigators to answer medical/biological questions using MPS data
- To enable scientists to make discoveries using resources available on campus
Course Overview

Day 1 – Sequence Generation

Day
Wednesday, August 1
Day 1: Sequence Generation

Time / Location
Session 1: 8:00AM to 12:00PM
Keish Conference Room 1289 CBRB

Session 2: 1:00PM to 5:00PM
Hartin Information Commons

Schedule
- “The Magic of MPS” (1 hour)
  Instructor: David Dimmock
- Course Overview – Background, Goals for the Course, and Glossary of Terms (30 minutes)
  Instructor: Elic Sheaar
- Next-Gen Sequencers Overview – Pros and Cons of each system; library preparation overview (50 minutes)
  Instructor: Kevin Knudtson
- Break 10:30 – 10:40
- Experimental Design – The importance of filtering: inheritance models, segregation analysis, limitations and possibilities, etc. (50 minutes)
  Instructor: Richard Smith
- Tour of the Genomics Core at the University of Iowa College of Medicine (30 minutes)
  Instructor: Kevin Knudtson
- Lunch 12:00 PM to 1:00 PM – EMR8 Atrium

Session 2:
- Introduction to Data Analysis with MPS Data Sets (30 minutes)
  Instructor: Ann Black-Ziegelmich
- Lab Session 1: Experimental Design for MPS Experiments – Using Galaxy to manipulate large data sets; creating a BED file for experimental design
  Instructors: Ann Black-Ziegelmich & Lab instructors

Course BBQ at Richard Smith’s Home, 11 Cherry Lane, 6-8PM
### Day 2 – Data Analysis

<table>
<thead>
<tr>
<th>Session 3: 8:00AM to 12:00PM</th>
<th>Kalch Conference Room 1289 CBRRB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality Check / Quality Assessment of MPS Data, Sequence Capture Assessment (50 minutes)</td>
<td>Andre Altman</td>
</tr>
<tr>
<td>Read Alignment Algorithms (50 minutes)</td>
<td>Tom Bair</td>
</tr>
<tr>
<td>Variant Calling Algorithms – SNVs and indels, in pooled and non-pooled data sets (1 hour)</td>
<td>Andre Altman</td>
</tr>
<tr>
<td>Detecting Copy Number Variants with MPS Data (1 hour)</td>
<td>Alex E. Nord</td>
</tr>
<tr>
<td>Lunch 12:00 PM to 1:00 PM – EMRRB Atrium</td>
<td></td>
</tr>
</tbody>
</table>

| Break 9:30 – 10:10 |

<table>
<thead>
<tr>
<th>Session 4: 1:00PM to 5:00PM</th>
<th>Hardin Information Commons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab Session 2: QC/QA of MPS data, Read Alignment, Variant Calling</td>
<td>Tom Bair &amp; Lab Instructors</td>
</tr>
</tbody>
</table>

### Day 3 – Data Interpretation

<table>
<thead>
<tr>
<th>Session 5: 8:00AM to 12:00PM</th>
<th>Kalch Conference Room 1289 CBRRB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Annotation and Filtering Overview (1 hour)</td>
<td>Alex E. Nord</td>
</tr>
<tr>
<td>Data management and data storage issues when dealing with NGS data sets (50 minutes)</td>
<td>Ben Rogers</td>
</tr>
<tr>
<td>Break 9:30 – 10:10</td>
<td></td>
</tr>
<tr>
<td>Practical considerations for grants and budgets (50 minutes)</td>
<td>Ben Heffron</td>
</tr>
<tr>
<td>Other Applications for MPS – RNASeq, Transcription, Methylation etc. (1 hour)</td>
<td>Tom Bair</td>
</tr>
</tbody>
</table>

| Lunch 12:00 PM to 1:00 PM – EMRRB Atrium |
| BREAKOUT Sessions – Q&A 1:00 PM to 2:00PM |

<table>
<thead>
<tr>
<th>Session 6: 2:00PM to 5:00PM</th>
<th>Hardin Information Commons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab Session 3: Annotating and Filtering Data Sets, Identifying Causal Variants, Exome Variant Data</td>
<td>Elliot Shearer &amp; Lab Instructors</td>
</tr>
</tbody>
</table>

5:00 PM Course adjourns
Tools for the Course

- **Galaxy**
  - A web-based platform
  - Makes command-line tools available to biologists
  - Flexible, sharable
  - Can be run from (almost) any computer

- **Helium Cluster**
  - A super-computing cluster
  - Cluster = a group of linked computers, working together thus in many respects forming a single computer
  - 3,508 processors, 1TB (1,000 Gb) of RAM

- **Course website:** [http://tinyurl.com/iowa-galaxy](http://tinyurl.com/iowa-galaxy)

- **Data:** from the OtoSCOPE platform

Sequences:
- Exons of non-syndromic genes and non-syndromic mimics
- 1,333 exons
- 521,647 bps
- Uses sequence capture and massively parallel sequencing

[http://morl-otoscope.org](http://morl-otoscope.org)

Molecular Otolaryngology and Renal Research Laboratories
Director: Richard J.H. Smith, M.D.
Outline

• Course Overview:
  § Who we are: Introduction of Course Organizers and Lecturers
  § Who you are: Background of Participants
  § Day-by-day
  § Tools for the course

• Introduction to Massively Parallel Sequencing
  § Whole Genome, Exome Sequencing, and Targeted Sequencing
  § Targeted Genomic Enrichment and Multiplexing
  § Examples

MPS Overview

• Massively parallel sequencing: a next-generation DNA sequencing technology that allows millions or billions of base-pairs to be sequenced simultaneously

<table>
<thead>
<tr>
<th>Technology</th>
<th>Traditional</th>
<th>Massively Parallel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Enrichment</td>
<td>Polymerase Chain Reaction (PCR)</td>
<td>Targeted Sequence Capture (solution or solid-phase)</td>
</tr>
<tr>
<td></td>
<td>- Single region (or multiplexed)</td>
<td>- Thousands of regions</td>
</tr>
<tr>
<td>DNA Sequencing</td>
<td>Chain termination</td>
<td>Sequence by synthesis</td>
</tr>
<tr>
<td></td>
<td>- $8,000 per megabase</td>
<td>- &lt;$1 per megabase</td>
</tr>
<tr>
<td></td>
<td>- 120,000 bp in 24 hrs</td>
<td>- 20 billion bp in 24 hours</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>Chromatograms</td>
<td>Sequencing Reads</td>
</tr>
<tr>
<td></td>
<td>-&gt; mapping, variant calling</td>
<td>- mapping, variant calling</td>
</tr>
</tbody>
</table>
Genomic DNA

Prepped DNA Library

PCR amplification 1

Adaptor Ligated Library

PCR amplification 2

Enriched Library for Sequencing

Illumina HiSeq

Sequence Capture

Multiplexing/Molecular Barcoding

- Only 1 sample per sequencing lane = sequencer output not maximized
- Molecular barcodes are integrated into sequencing adaptors during library preparation
Multiplexing

Genomic DNA

Barcoded adaptor-ligated library

Pooled DNA

Illumina HiSeq

Multiplexed Sequence Capture

Genomic DNA

PCR amplification 1

Hybridization

PCR Amplification 2

Barcoded adaptor-ligated library

Illumina HiSeq
The Scope of MPS

the human genome

- 1 square = 10,000,000 bp

The exome: Sequencing every exon of every gene in the genome

50,000,000 bp

Targeted Capture: OtoSCOPE®

500,000 bp

Exome Sequencing – A Bridge

<table>
<thead>
<tr>
<th>Method</th>
<th>OtoSCOPE®</th>
<th>Whole Exome</th>
<th>Whole Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sequence Capture and Sequencing</td>
<td>Sequence Capture and Sequencing</td>
<td>Sequencing</td>
</tr>
<tr>
<td>Base pairs sequenced</td>
<td>~500,000 bp</td>
<td>~50,000,000 bp</td>
<td>3,200,000,000 bp</td>
</tr>
<tr>
<td>Variants discovered</td>
<td>300</td>
<td>20,000</td>
<td>362,000</td>
</tr>
<tr>
<td>Cost (reagents only)</td>
<td>$350</td>
<td>$1,000</td>
<td>$16,413</td>
</tr>
</tbody>
</table>

Fisher et al, Genome Biology, 2011
Depristo et al, Nature Genetics, 2011
Examples

Nature Genetics 2009

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng1,2,3,4, Kari I Buckingham1,2,3, Choli Lee1,4, Abigail W Bigham2, Holly K Talbot4,5, Karie M Dent3, Chad D Huff2, Paul T Shannon5,6, Ethelynn Wang-Irbe5, Deborah A Nickerson4,5, Jay Shendure1,7,8 and Michael J Bamshad1,2,3

AJHG 2010

Targeted Capture and Next-Generation Sequencing Identifies C9orf75, Encoding Taperin, as the Mutated Gene in Nonsyndromic Deafness DFNB79

Atteeq Ur Rehman1, J. Robert J. Morell, Inna A. Belyaevskaya, Shahid Y. Khan, Etish T. Roger, Mohsin Shahraz, Zubair M. Ahmed, Salma Riazuddin, Shaheen N. Khan, Sheikh Riazuddin, and Thomas B. Friedman1,2

PNAS 2010

Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing

Tom Walsh6, Ming K. Lee5, Ulhas Gadad6, Anna M. Thornton6, Sunday M. Stroy1, Christopher Fareed1, Alex S. Norad6, Jessica A. Medoff6, Elizabeth M. Yandell5, and Mary Clare King6

Genet Med, 2010.© 2010 Massachusetts Medical Society. All rights reserved.

Table 1 Direct identification of the gene for a mendelian disorder by exome resequencing

<table>
<thead>
<tr>
<th>Filter</th>
<th>Kindred 1 A</th>
<th>Kindred 1 B</th>
<th>Kindred 1 (A+B)</th>
<th>Kindred 1+2</th>
<th>Kindred 1+2</th>
<th>Kindred 1+2+3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dominant</td>
<td>Recassive</td>
<td>Dominant</td>
<td>Recassive</td>
<td>Dominant</td>
<td>Recassive</td>
</tr>
<tr>
<td>NS/SNP</td>
<td>4,870</td>
<td>2,803</td>
<td>4,687</td>
<td>2,809</td>
<td>3,960</td>
<td>3,999</td>
</tr>
<tr>
<td>Not in dSNP129</td>
<td>641</td>
<td>102</td>
<td>647</td>
<td>114</td>
<td>369</td>
<td>53</td>
</tr>
<tr>
<td>Not in Haplo8</td>
<td>898</td>
<td>123</td>
<td>923</td>
<td>128</td>
<td>506</td>
<td>46</td>
</tr>
<tr>
<td>Not in either</td>
<td>406</td>
<td>31</td>
<td>434</td>
<td>33</td>
<td>228</td>
<td>5</td>
</tr>
<tr>
<td>Predicted damaging</td>
<td>204</td>
<td>6</td>
<td>204</td>
<td>12</td>
<td>83</td>
<td>1</td>
</tr>
</tbody>
</table>
• Combined traditional (linkage analysis) with MPS (targeted sequence capture)
• Shows possible pitfalls with respect to QC/QA

---

• The first study using a targeted genomic enrichment panel for diagnosis
• Used CNV calling from MPS data successfully
Thank You and Have a Great Course!