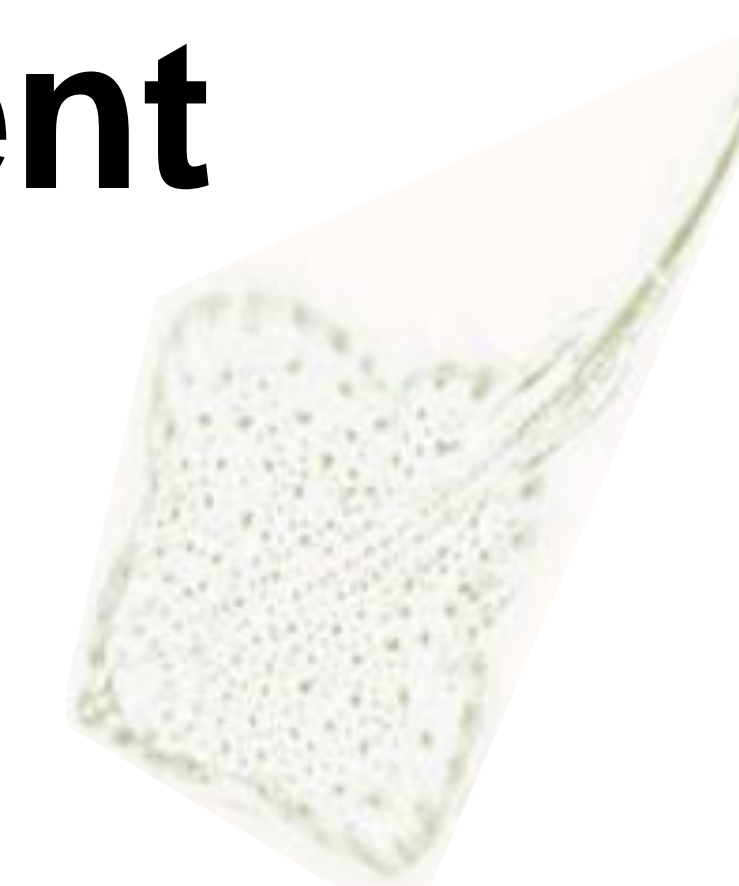




# Skate Genome Project: a Cyber-Enabled Research and Workforce Development

Qinghua Wang<sup>1</sup>, Benjamin King<sup>2</sup>, Cecilia Arighi<sup>1</sup>, Hongzhan Huang<sup>1</sup>, Mihailo Kaplarevic<sup>1</sup>, Bruce Kingham<sup>1</sup>, James Vincent<sup>3</sup>, Marc Farnum Rendino<sup>3</sup>, Randall D. Dahn<sup>2</sup>, Carolyn Mattingly<sup>2</sup>, Clinton Chichester<sup>4</sup>, Joanna Fueyo<sup>4</sup>, William Kelley Thomas<sup>5</sup> and Cathy H. Wu<sup>1</sup>

<sup>1</sup>Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE; <sup>2</sup>Mount Desert Island Biological Laboratory, Salisbury Cove, ME; <sup>3</sup>University of Vermont, Burlington, VT; <sup>4</sup>University of Rhode Island, Kingston, RI; <sup>5</sup>University of New Hampshire, Durham, NH



## Background and Objective

In 2006, the North East IDeA states (VT, NH, ME, RI, DE) launched an initiative now called the North East Cyberinfrastructure Consortium (**NECC**). Two related collaborative efforts were initiated: to identify and promote the shared use of research facilities across the region, and assess and address cyberinfrastructure needs. The Little Skate Genome Sequencing & Annotation Project provides a model for highly collaborative approaches to use specialized resources and expertise in an integrated process.

## Introduction to the Skate Genome Project

The little skate, *Leucoraja erinacea*, as a model organism for vertebrate phylogeny with a comparative approach, i.e. to exploit natural diversity and to infer and understand ancestral mechanisms, including:

- Model for renal physiology
- Model for limb regeneration: understand lineage-specific variation

## Scientific Significance:

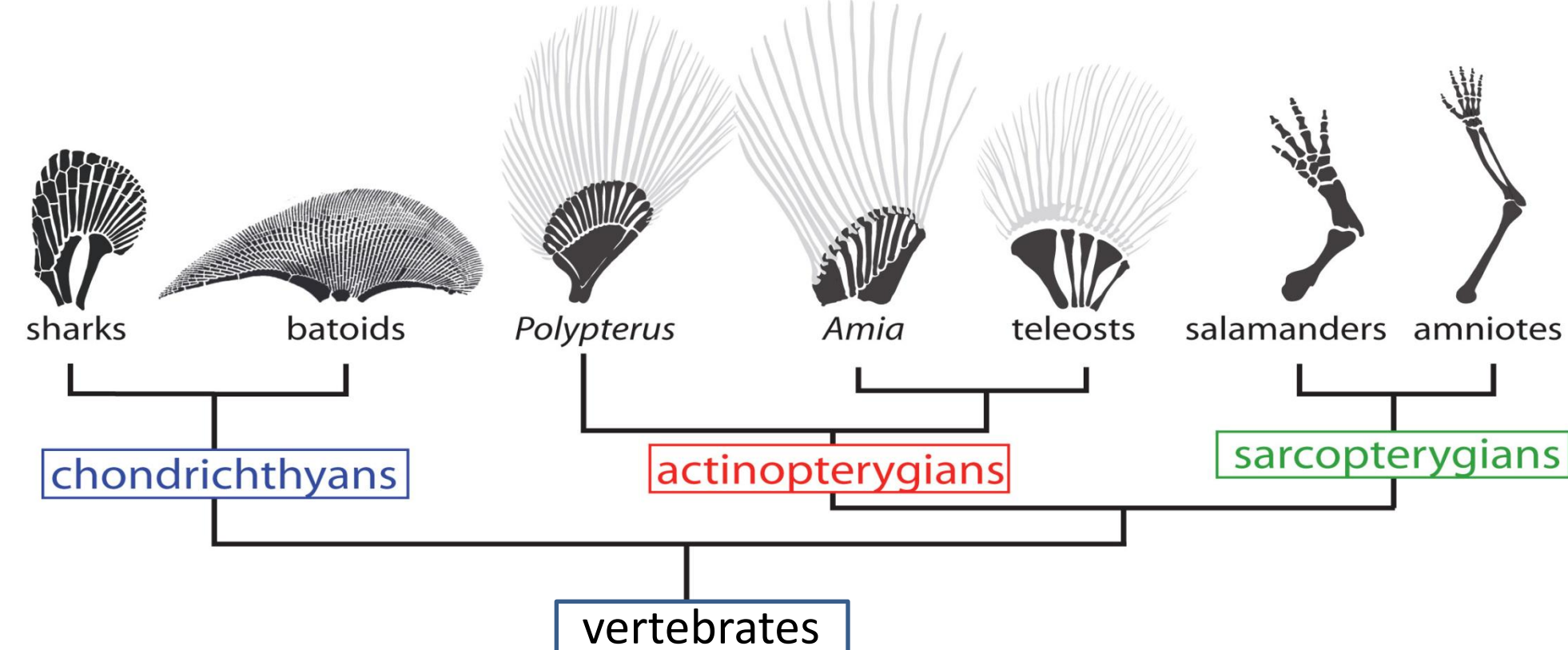
*The skate genome will significantly advance comparative functional genomics research.*

The skate is a chondrichthyan fish that appeared approximately 450 million years ago, closely following the initial appearance of jaws and paired appendages in the vertebrate fossil record.

## Skate Genome

Haploid Genome Size:

- 49 chromosomes
- Estimated size is 3.42 Gbp



## Skate Sequences Prior to Project

Genomic sequences:

HOXA cluster (134kb)

Expressed Sequence Tags:

- 31,116 ESTs from 4 cDNA libraries:
- 14,726 ESTs from normalized mixed tissue library
- 6,016 ESTs from normalized liver library
- 5,600 ESTs from embryonic library
- 4,825 ESTs from an embryonic cell line library

192 non-EST sequences in GenBank

Ongoing at MDIBL:

- Transcriptome sequencing
- microRNA sequencing



## Genomic Resources

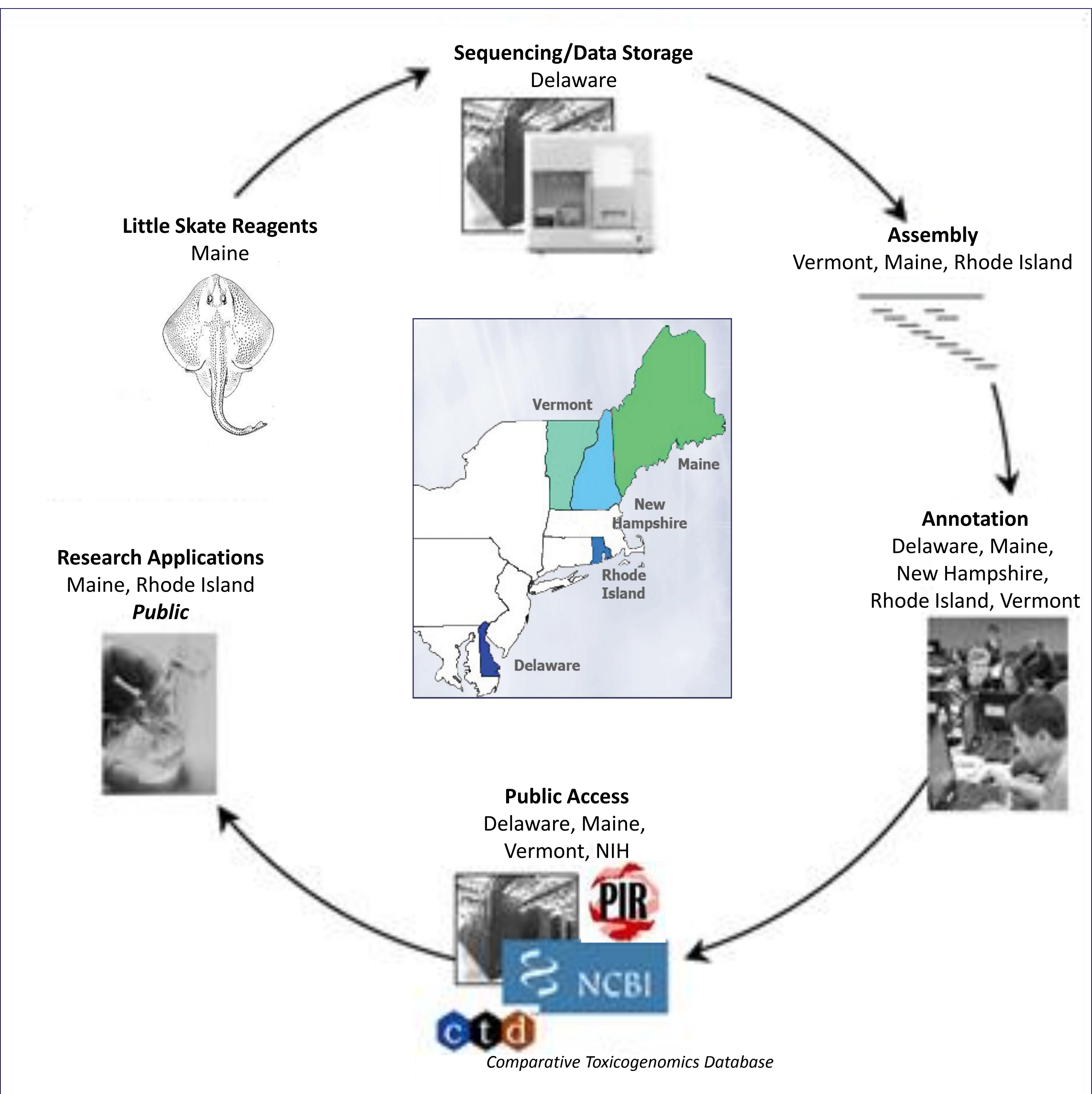
4x Bacterial Artificial Chromosome library from MDIBL

## Results

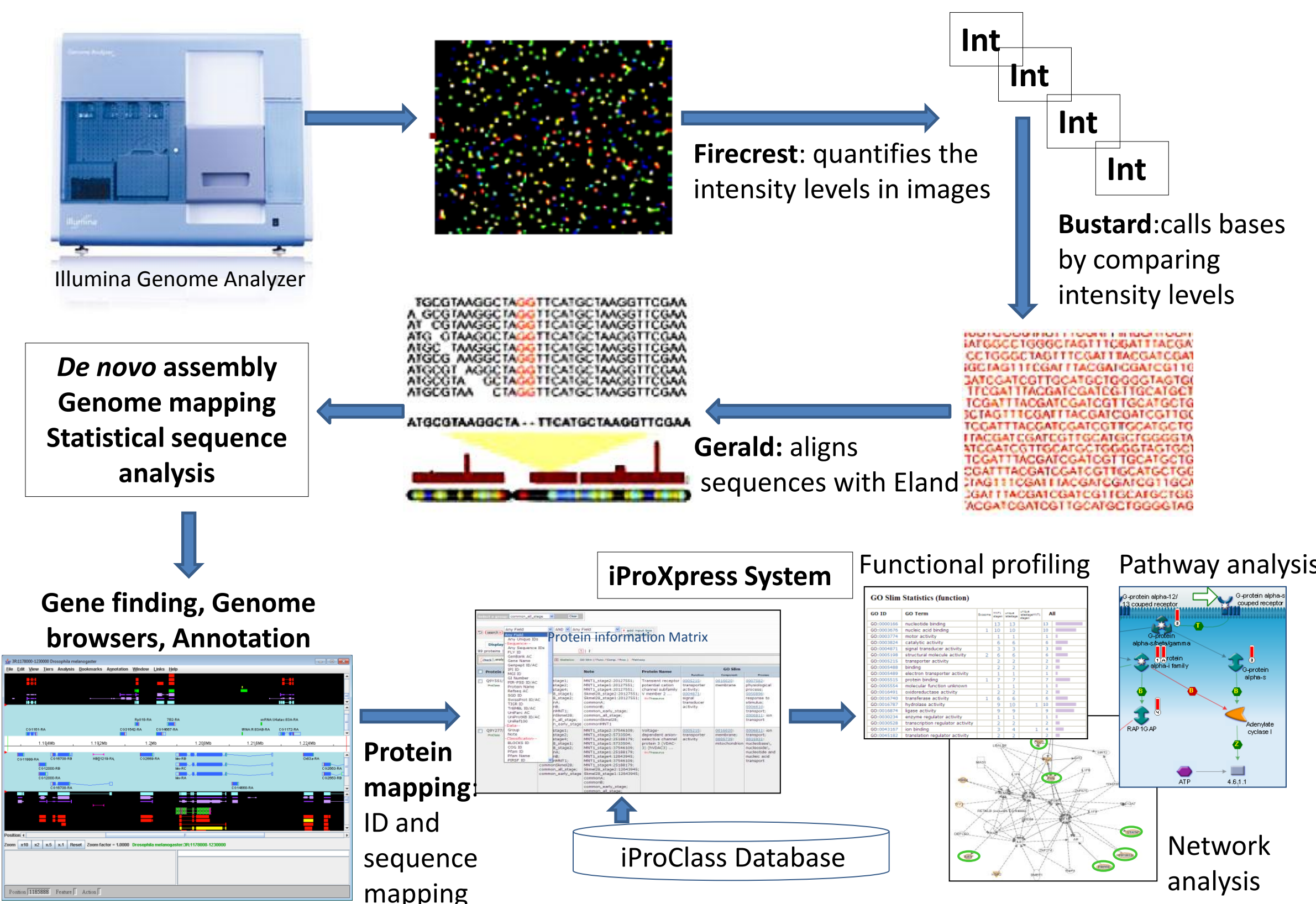
Preliminary sequencing results have been collected and sequence assembly and analysis pipeline has been built. One-week long genome annotation workshop was organized, which covered lectures, tutorials, annotation exercises and group discussion. In addition, online distributed annotation jamborees will be adopted monthly via videoconferencing and web-based annotation interface.

## Collaborative Use of Specialized Resources & Expertise

- Little Skate Reagents (Maine)
- Next-Generation Sequencing: UD DNA Sequencing & Genotyping Center (DE)
- Sequence Assembly: Vermont Genetics Network (VT) with ME, RI
- Sequence Analysis/Annotation: pipeline at DE, ME, RI, NH and VT
- Storage & Access Of Sequence/Annotation Data: UD shared data center (DE)
- Public Dissemination: UD CBCB (DE), MDIBL (ME), *skatebase.org* (VT), NIH



## Analysis Pipeline for Next-Gen Sequencing Data



## Bioinformatics Workforce Development

### Collaborative and Distributed Training

- Integrated with research on little skate genome sequencing and annotation
- Genome annotation workshops (DE, ME): one-week long semiannually, covering lectures, tutorials, annotation exercises
- Online distributed annotation marathons: one-day long, monthly via videoconferencing and web interface for extended training/genome annotation
- Coupling training with annotation helps improve understanding of the underlying bioinformatics approaches and methods
- Participants: Trainees from NECC institutions, regional minority/UG institutions

### Genome Annotation Workshops and Jamborees

- May 24-28, 2010, UD CBCB (DE)
- October 19-22, 2010, MDIBL (ME)
- May 23-27, 2011, UD CBCB (DE)

### First Skate Genome Annotation Workshop

**Participants:** 27 students and faculty from 11 universities and colleges

**Instructors:** 11 instructors from 7 universities and institutes

**Lectures/Tutorials & Hands-on Exercises on the Following Topics:**

- Genome sequencing using Next-Gen
- Genome Assembly and Analysis
- Introduction to sequence, gene and protein resources
- Sequence similarity searching
- Genome sequence analysis, Gene finding, Genome browsers
- Genome annotation platform
- Protein Bioinformatics: Protein database resources & text searches, sequence and structure-based analysis, literature mining, biomedical ontologies

### Genome Annotation Jamboree

Visit to the Next-Gen Sequencing Lab at UD



## Acknowledgements

Funding: Delaware INBRE 3P20RR016472-09S2



The University of Vermont

