Preview of Award 1062301 - Annual Project Report

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1062301
Project Title:	ABI Innovation: Barcoding-Free Multiplexing: Leveraging Combinatorial Pooling for High-Throughput Sequencing
PD/PI Name:	Stefano Lonardi, Principal Investigator Timothy J Close, Co-Principal Investigator
Submitting Official (if other than PD\PI):	Stefano Lonardi Principal Investigator
Submission Date:	04/19/2013
Recipient Organization:	University of California-Riverside
Project/Grant Period:	05/01/2011 - 04/30/2014
Reporting Period:	05/01/2012 - 04/30/2013
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Stefano Lonardi

Accomplishments

*** What are the major goals of the project?**

The research plan for this NSF award is articulated around a novel sequencing protocol that combines next-generation sequencing instruments and 'smart' pooling. The proposed protocol hinges on the computational component, which deals with the preprocessing of short reads and post-processing of contigs. The development of this computational component was the focus of the our effort in Year 1. In Year 2 (this year) we applied the new protocol on a BAC library for barley (*Hordeum vulgare*) and cowpea (*Vigna unguiculata*).

The main steps of our combinatorial sequencing method are summarized next. More details can be found in the original grant proposal.

- A. Obtain a BAC library for the target organism
- B. Select gene-enriched BACs from the library (optional)
- C. Fingerprint BACs and build a physical map
- D. Select a minimum tiling path (*MTP*) from the physical map
- E. Pool the MTP BACs according to the shifted transversal design
- F. Sequence the DNA in each pool, trim/clean sequenced reads
- G. Clean and error-correct reads and generate unireads by greedy assembly
- H. Assign unireads to BAC clones (*deconvolution*)
- I. Assemble reads BAC-by-BAC using a short-read assembler
- J. Merge BAC assemblies (guided by the physical map)

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities:

Accomplishments on steps A-E for the cowpea genome were reported in Year 1 report. For barley, we carried out a similar strategy. The outcome of the effort on the barley genome was reported (among other resources) in the paper appeared in *Nature* (Nov 2012) on the first draft of the barley genome.

In Year 2, we focused on steps F, G, H and I. More specifically:

- we sequenced two sets of 2,197 cowpea BAC clones using the Illumina HiSeq 2000 (step F)
- we sequenced seven sets of 2,197 barley BAC clones using the Illumina HiSeq 2000 (step F)
- we carried out decoding of the reads for both cowpea and barley using our software tool Hashfilter (available in the public domain and described in the *PLoS Computational Biology* manuscript) (step H)
- we assembled each cowpea and barley BAC (step I) and made the sequence of this clones available to the community

Specific Objectives:

- We have spend a considerable amount of time developing the software tool to decode the reads to the correct BAC: we used simulated data for the rice genome to make sure that the tool was accurate and efficient
- Since we are dealing with several TB of sequenced data, we have invested in a high-memory file server that has been critical in carrying out this research
- We have gained considerable experience in dealing with the data produced by the Illumina HiSeq 2000, and we are confident that we were able to 'squeeze' the most out of these datasets

Significant Results:

- Our software HashFilter is currently capable of consistently decoding 70-80% of the reads for all barley and cowpea datasets (intially the fraction was as low as 25%)
- On simulated data (rice) 99.57% of the deconvoluted reads are assigned to the correct BAC
- The pipeline to carry out BAC-by-BAC assemblies uses Velvet, and the average N50 is very high, often over 20K bases (i.e., BAC assembly have very high quality)

Key outcomes or Other achievements:

- The paper in *PLoS Comp Bio* describing the protocol has been out only a few weeks, and it has already generated significant attention from the community (for instance, I have been invited to give a webinar to explain the steps involved in the pooling and the data processing from a large website devoted to plant community resources)

*** What opportunities for training and professional development has the project provided?**

This project is allowing us to train two PhD students (Computer Science) one of which is a female. One MS student (Computer Science), and one undergraduate student (Computer Science) were involved in Year 1. All students have been trained in the domain of computational biology. Specifically:

- PhD student Denise Duma has been working on the simulation of the protocol on the rice genome and helped in the design of the deconvolution algorithm. She is currently investigating a method to carry out the error-correction step.

- PhD student Hamid Seyed Mirebrahim has been working on the problem of integrating BAC assemblies with the whole genome shotgun assembly.
- MS student Burair Alsaihati worked on earlier version of the tool to compute prefix-suffix overlaps between all pairs of reads. He graduated in the Summer of 2011, and currently work in Joint Center for Genomic Research in his country (Saudi Arabia).
- Undergraduate student Matt Alpert has been working on the BAC-by-BAC assemblies of the rice data. Two summers ago, I requested and obtained additional funding as an REU to support him. He is currently enrolled in the PhD program in my department and continues to collaborate with us on this project.
- Post-doc Francesca Cordero (University of Torino, Italy) and post-doc Marco Beccuti (University of Torino, Italy) visited my lab during the summer of 2011. In collaboration with Francesca, Marco and another faculty in my department (Prof. G. Ciardo) we developed and implemented the software HashFilter (descrived in the *PLoS Computational Biology* manuscript).

*** How have the results been disseminated to communities of interest?**

- Results of our research was published in *Nature* (2012) and *PLoS Computational Biology* (2013)
- Software Hashfilter is available in the public domain at http://www.cs.ucr.edu/~stelo/hashfilter/
- Raw sequence reads for barley have been deposited in NCBI Sequence Read Archive
- Barley assembled BACS are available to the community via http://www.harvest-web.org /utilmenu.wc?job=RTRVFORM&db=MOREX_HV3_9
- Cowpea assembled BACs are available to the community via http://www.harvest-web.org /utilmenu.wc?job=RTRVFORM&db=COWPEA_BAC
- PI Lonardi gave a talk entitled "Combinatorial Pooling Enables Selective Sequencing of the Barley Gene Space" in June 2012 at the workshop "High-Throughput Sequencing and Other Methods: from Technology to Discovery", University of California, Irvine, CA
- PI Lonardi gave a talk entitled "Combinatorial Pooling Enables Selective Sequencing of the Barley Gene Space" in Febrary 2012 at the workshop "IMA workshop Group Testing Designs, Algorithms, and Applications to Biology", University of Minnesota, Minneapolis
- PI Lonardi gave a talk entited "Combinatorial Pooling Enables Selective Sequencing of the Barley Gene Space" in April 2012 at University of California, Riverside, Department of Computer Science and Engineering

*** What do you plan to do during the next reporting period to accomplish the goals?**

Goals for Year 3:

- Analyze the cowpea BAC clones (i.e., gene finding, functional annotations, compare with common bean, etc.), integrate them with the whole genome shotgun assembly, the write a manuscript that summarizes the finding on this genome
- Analyze the barley BAC clones (i.e., gene finding, functional annotations, etc.) then write a manuscript that summarizes the finding on this genome
- Design an improved decoding algorithm that is more accurate than HashFilter
- Develop a read error-correcting tool that exploits the pooling design

---

Products

Journals

N. Stein, ... K. Madishetty, M. Moscou, P. Bhat, S. Wannamaker, T. Close, Y. Ma, D. Duma, F. Cordero, G. Ciardo, M. Beccuti, M. Alpert, S. Lonardi, ..., A. Zuccolo, F. Cattonaro, M. Morgante, S. Scalabrin, S. Radovic, V. Vendramin, J. Poland, R. Wise (11/29/12). A physical, genetic and functional sequence assembly of the barley genome. *Nature*. 490 (7422), 711-716.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.1038/nature11543

S. Lonardi, D. Duma, M. Alpert, F. Cordero, M. Beccuti, P. R. Bhat, Y. Wu, G. Ciardo, B. Alsaihati, Y. Ma, S. Wanamaker, J. Resnik, S. Bozdag, M-C. Luo, T. J. Close (4/5/13). Combinatorial Pooling Enables Selective Sequencing of the Barley Gene Space. *PLoS Computational Biology*. 9 (4), e1003010.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: doi:10.1371/journal.pcbi.1003010

Books

Book Chapters

Thesis/Dissertations

Conference Papers and Presentations

Other Publications

Technologies or Techniques
Nothing to report.

Patents
Nothing to report.

Inventions
Nothing to report.

Licenses
Nothing to report.

Websites
Nothing to report.

Other Products

Product Type:	Software or Netware
Description:	Software for decoding reads to the source BAC clone, when DNA samples are pooled according the shifted transversal design: http://www.cs.ucr.edu/~stelo/hashfilter/ Associated with the PLoS Comp Biology paper.
Other:	Databases
Product Type:	Barley BAC assemblies from
Description:	http://www.harvest-web.org/utilmenu.wc?job=RTRVFORM&db=MOREX_HV3_9 Cowpea BAC assemblies from http://www.harvest-web.org/utilmenu.wc?job=RTRVFORM&db=COWPEA_BAC

Other:

Participants

Research Experience for Undergraduates (REU) funding

How many REU applications were received during this reporting period? 0

How many REU applicants were selected and agreed to participate during this reporting period? 0

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Denisa Duma	Graduate Student (research assistant)	6
Stefano Lonardi	PD/PI	2
Timothy J Close	Co PD/PI	2
Seyed Mirebrahim	Graduate Student (research assistant)	6
Steve Wanamaker	Technician	4

What other organizations have been involved as partners?
Nothing to report.

Have other collaborators or contacts been involved? N

Impacts

What is the impact on the development of the principal discipline(s) of the project?

The objective of this project is to facilitate the sequencing of large, highly repetitive genomes like the genome of barley (5.3 GB -- twice the size of human) and cowpea (650-700 MB). The availability of this new sequencing protocol and the availability of these two genome constitute the major impact of this project.

What is the impact on other disciplines?

Cowpea also known as black-eyed pea, is a primary source of protein in the human diet in Sub-Saharan Africa, where it is grown for its foliage, and fresh and dry grains. Outside Africa, cowpea is grown in parts of Asia, Latin America, Southeastern USA and California. Despite its relevance to agriculture in the developing world, cowpea has received scant attention relative to other crops of major global significance. We believe that producing the primary DNA sequence of this important crop will have a significant scientific impact as well enable scientists to select and engineering specific traits that would help ensure high yields from sustainable agriculture into the future.

What is the impact on the development of human resources?

Training of graduate and undergraduate students in computational biology and genomics.

What is the impact on physical resources that form infrastructure?

N/A

What is the impact on institutional resources that form infrastructure?

N/A

What is the impact on information resources that form infrastructure?

N/A

What is the impact on technology transfer?

We are currently exploring whether our technology might have a market (commercial) value.

What is the impact on society beyond science and technology?

The long-term impact of sequencing another organism can be profound once the sequence becomes a routine component of practical problem solving, but in addition the process of sequencing an organism provides precious experience for young scientists who then become our continuity into the future. The organisms that have been sequenced so far represent a minuscule proportion of those which live on our planet, many of which we depend on for our food or against which we must defend to sustain human society. In the plant world for example, the outcome of sequencing an organism could lead to selection and engineering of specific traits that would help ensure high yields from sustainable agriculture into the future. The current food crisis has been fueled by an increase in food demands by developing countries, the spike in oil prices with impact on the cost of fertilizers, disruptions due to climate change and the push to produce biofuels. A global food crisis now severely effects at least 36 mainly developing countries and at least one billion people.

Changes

Changes in approach and reason for change

N/A

Actual or Anticipated problems or delays and actions or plans to resolve them

N/A

Changes that have a significant impact on expenditures

N/A

Significant changes in use or care of human subjects

N/A

Significant changes in use or care of vertebrate animals

N/A

Significant changes in use or care of biohazards

N/A

Special Requirements

Responses to any special reporting requirements specified in the award terms and conditions, as well as any award specific reporting requirements.

N/A