Annual Report for Period: 05/2011 - 04/2012
Submitted on: 02/02/2012
Principal Investigator: Lonardi, Stefano .
Award ID: 1062301
Organization: U of Cal Riverside
Submitted By:
Lonardi, Stefano - Principal Investigator
Title:
ABI Innovation: Barcoding-Free Multiplexing: Leveraging Combinatorial Pooling for High-Throughput Sequencing

Project Participants

Senior Personnel

Name: Lonardi, Stefano
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Close, Timothy
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Ciardo, Gianfranco
Worked for more than 160 Hours: Yes
Contribution to Project:
collaborator, no support

Post-doc

Name: Cordero, Francesca
Worked for more than 160 Hours: Yes
Contribution to Project:
Visiting post-doc for the summer of 2011 (see attached document)

Name: Beccuti, Marco
Worked for more than 160 Hours: Yes
Contribution to Project:
Visiting post-doc for the summer of 2011 (see attached document)

Graduate Student

Name: Duma, Denisa
Worked for more than 160 Hours: Yes
Contribution to Project:
Supported on this grant (see attached document)

Name: Alsaihati, Burair
Worked for more than 160 Hours: Yes
Contribution to Project:
MS student, now graduated (see attached document)

Undergraduate Student

Name: Alpert, Matthew
Worked for more than 160 Hours: Yes
Contribution to Project:

Supported for the summer via a REU (see attached document)

Technician, Programmer

Name: Wanamaker, Steve
Worked for more than 160 Hours: Yes
Contribution to Project:
Sys administrator and programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Activities and Findings

Research and Education Activities:
see attached document

Findings:
see attached document

Training and Development:
see attached document

Outreach Activities:
see attached document

Journal Publications

Books or Other One-time Publications

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:
see attached document

Contributions to Other Disciplines:
see attached document

Contributions to Human Resource Development:

see attached document

Contributions to Resources for Research and Education:
see attached document

Contributions Beyond Science and Engineering:
see attached document

Conference Proceedings

Special Requirements

Special reporting requirements: None
Change in Objectives or Scope: None
Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Organizational Partners
Any Journal
Any Book
Any Web/Internet Site
Any Product
Any Conference

1 Introduction

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. We proposed to apply the new protocol on a BAC library for cowpea (*Vigna unguiculata*).

The main steps of our *combinatorial sequencing* method are summarized next and illustrated in Figure 1. More details can be found in the original 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 [4, 2]

E. Pool the MTP BACs according to the shifted transversal design [10]

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)

1.1 Progress on Steps A-E: Building MTP Pools for Cowpea

We started from a 17X depth of coverage cowpea BAC library containing about 60,000 BACs from the African breeding genotype IT97K-499-35 with an average insert size of 150 kb (step **A**). Cowpea BACs were fingerprinted at UC Davis using high information content fingerprinting(HICF) [3, 7, 8]. Based on the fingerprinting data, a physical map was produced from 43,717 BACs (see `http://phymap.ucdavis.edu/cowpea/`) with a depth of 11x genome coverage (step **C**) [1, 12], and a minimal tiling path (MTP) comprised of 4394 clones was derived (step **D**) [2]. The set of MTP clones was split in two sets of $N = 2197$ BACs, each of which was pooled according to the shifted transversal design (step **E**).

Recall that we pooled these two subsets of 2,197 BACs according to the shifted transversal design [10]. This pooling design is defined by three parameters (P, L, Γ). Taking into consideration the format of the standard 96-well plate and the need for a 3-decodable pooling design for MTP BACs[1], we chose parameters

[1] If the MTP was truly a set of minimally overlapping clones, a two-decodable pooling would be sufficient, but a three-decodable pooling gives additional protection against errors.

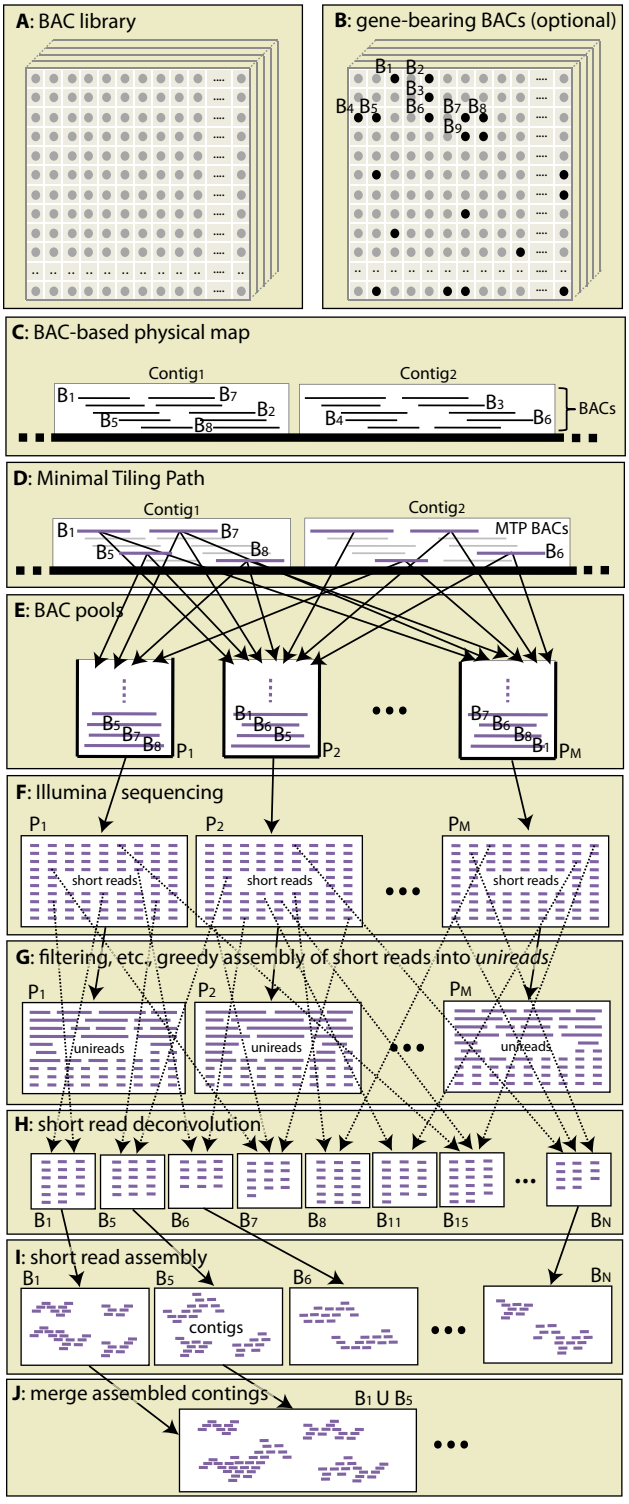

Figure 1: The sequencing protocol originally proposed in the grant application.

$P = 13$, $L = 7$ and $\Gamma = 2$, so that $P^{\Gamma+1} = 2{,}197$ and $\lfloor (L-1)/\Gamma \rfloor = 3$. Each of the $L = 7$ layers consisted of $P = 13$ pools, for a total of 91 BAC pools, which left some space for a few control DNA samples on the 96-well plate. In this pooling design, each BAC is contained in $L = 7$ pools and each pool contains $P^{\Gamma} = 169$ clones. We call the set of L pools to which a BAC is assigned, the *BAC signature*. Any two BAC signatures can share at most $\Gamma = 2$ pools.

1.2 Progress on Steps F-G: Sequencing and Processing Paired-end Reads

The sequencing of the two sets of 2,197 cowpea BAC pools is currently in progress. We should obtain the sequence data from the Illumina HiSeq 2000 installed at UC Riverside in the next month or two. In the meantime we have designed and tested our tool on simulated data on the rice genome (see Section 2).

The research questions in the original proposals were the following.

- G1: Investigate error-correcting tools for short reads, possibly develop novel algorithms/tools
- G2: Develop an algorithm/tool to compute unireads: the problem requires to determine approximate prefix-suffix matching between any two pairs of reads in a sequenced pool (composed of 2-3 million reads); trying to match all pairs is computationally unfeasible

Regarding objective G1, we are currently working on an implementation of an error-correcting tool specifically designed to take advantage of the structure of the pooling. The main idea is that if a read does not appear in the set of pools corresponding to one of the BAC signature, then either it is a repetitive read or it contains a sequencing error. A k-mer analysis of those reads should allow us to be able to pinpoint the locations of sequencing errors and correct them.

Regarding question G2, we abandoned the idea of generating *unireads* entirely. We have developed a method to compute the approximate prefix-suffix overlap that is sufficiently time-efficient and space-efficient on the original reads (see Section 1.3).

1.3 Progress on Step H: Deconvoluting Paired-end Reads to BACs

The research questions in the original proposals were the following.

- H1: Design an algorithm/tool to compute the signature set O_r for each read r; since each pool could generate a few million reads, it is not feasible to try to approximately match each read r against all the others
- H2: Once O_r is available for all reads, design an algorithm/tool to deconvolute each read to one or more BACs; note that the deconvolution process has to the repeated hundreds of millions of times; if implemented as described above it will be too time consuming

The computational problems described here as H1 and H2 have been the main focus of the first year of this NSF award. We spent a considerable amount of time to devise a time- and memory-efficient method to carry out the deconvolution, which is summarized next.

Let $M = P \times L$ be the total number of pools and R_i be the set of reads obtained by sequencing pool i, for all $i \in [1, M]$. Since each pool set R_i can contain tens of million of reads the computation of the read signature computation can be extremely computationally intensive if implemented naively. In our first attempt, we computed pairwise prefix-suffix approximate overlaps between the reads in all $M(M-1)/2$ pairs of pools. We used quite sophisticated implementation by Välimäki *et al.* [11] and Simpson *et al.* [9] using the FM-index [5, 6] but they turned out to be still too inefficient for our purposes. While the rice dataset took about a week of computation, processing the reads in a more repetitive datasets would have taken months on a multi-core architecture.

We devised an alternative strategy that skips entirely the prefix-suffix overlap between all pairs of reads, but instead uses the k-mer content of each read to identify which set of pools it belongs. For each set R_i we first compute the frequency $count_i$ of all its distinct k-mers. Specifically, for each k-mer $w \in R_i$, $count_i(w) = c$ if w or its reverse complement occurs exactly c times in R_i. These counts are stored in a hash table. For each distinct k-mer w the table stores a frequency vector of M numbers, namely $[count_1(w), count_2(w), \ldots, count_M(w)]$. Once the table is built, we process each read again as follows. Given a read r in R_i, we fetch the frequency vectors for all its k-mers from the hash table. If the number of positive counts for a k-mer is not in the set $A = \{L, 2L - \Gamma, \ldots, 2L, 3L - 2\Gamma, \ldots, 3L\}$, then the minimum number of lowest counts are removed from the vector to make the number of positive counts equal to one of the elements in A. Recall that by construction each BAC is assigned to L pools, thus the *signature* of a BAC is a set of L numbers in the range $[1, M]$. Moreover, two BAC signatures cannot share more than Γ pools (see Theorem I in [10]). Each k-mer frequency vector is "binarized", then matched against the BAC signatures: if no good match exists, its frequency vector is discarded. At the end of this process, the frequency vectors that correspond to a valid BAC signature are combined to form the *signature* of read r.

This algorithm has been implemented in the multi-threaded tool HASHFILTER which has been used to run the simulations on the rice genome reported in Section 2.

1.4 Progress on Steps I-J: Assembling and Merging Partial Assemblies

The research questions in the original proposals were the following.

- I1: Carry out an extensive benchmarking of existing short-read assembly tools on synthetic data; explore the sensitivity of the method to read length, number of reads, sequencing error rate, deconvolution error rate, etc.
- J1: Design an algorithm/tool to produce the final assembly from partial BAC-by-BAC assemblies that takes advantage of the physical map
- J2: Compare the BAC-by-BAC cowpea assembly with the whole-genome cowpea shotgun assembly; eventually merge the two assemblies

We have gained considerable experience on I1, by running several assembly tools (VELVET, SOAPDENOVO and ABYSS) on the simulated data for rice (see Section 2). We have not yet worked on objective J1, but we have a clear idea on how to carry out this step.

Regarding alternative cowpea assemblies (J2), a team composed by the PIs, Scott Jackson (Purdue University) and Greg May (National Center for Genome Resources) has sequenced the whole cowpea genome using a whole-genome shotgun approach using a single run of the Illumina IIx ($\approx$125 bases, paired-end reads). The assembly had a rather shallow sequencing depth and the resulting contig do not provide sufficient information about genes and other non-repetitive regions. Still, the comparison between the assemblies obtained by our protocol and the whole-genome shotgun will be very informative and will constitute a validation of our method. Eventually the two independently obtained assemblies will be merged.

2 Experimental Results on the Rice Genome

The physical map for *Oryza sativa* was assembled from 22,474 BACs fingerprinted at AGCoL, and contained 1,937 contigs and 1,290 singletons. From this map, we selected only BACs whose sequence could be uniquely mapped to the rice genome. We computed an MTP of this smaller map using our tool FMTP [2].

The resulting MTP contained 3,827 BACs with an average length of ≈ 150 kb, and spanned 91% of the rice genome (which is ≈ 390 Mb).

We pooled *in silico* a subset of 2,197 BACs from the set above according to the shifted transversal design [10]. This pooling design is defined by three parameters (P, L, Γ). We used the same pooling parameters discussed in Section 1.1 ($P = 13$, $L = 7$ and $\Gamma = 2$).

The 91 resulting rice BAC pools were "sequenced" *in silico* by generating 10^6 paired-end reads of 104 bases with an insert size of 327 bases, and 1% sequencing error distributed uniformly along the read. A total of 208M usable bases gave an expected ≈ 8x sequencing depth for a BAC in a pool. As each BAC is present in seven pools, this is an expected ≈ 56x combined coverage.

The 91 read pools were processed for deconvolution via k-mers analysis ($k = 26$ in our experiments). The computation was relatively quick, but required a significant amount of memory. The construction of the hash table required about 120 GB of RAM and 164 minutes running on one core of a Dell PowerEdge T710 server (dual Intel Xeon X5660 2.8Ghz, 12 cores, 144 Gb RAM). The deconvolution phase took 33 minutes on 10 cores; sorting the reads into 2,197 files took 22 minutes (one core).

Figure 2-(a) illustrates the distribution of signature sizes for all the distinct k-mers in the rice dataset. Observe that the distribution has clear peaks around $L = 7$, around the interval $[2L - \Gamma, 2L] = [12, 14]$ and the interval $[3L - 3\Gamma, 3L] = [15, 21]$. These peaks correspond to k-mers originating from one, two, and three overlapping BACs, respectively. We also have a rather large number of k-mers appearing in 1–5 pools. For a k-mer to have fewer than $L = 7$ occurrences, sequencing errors must have occurred (assuming the sequencing depth to be sufficient and excluding technical errors with BACs). Figure 2-(b) shows the distribution of signature sizes for all the reads in the rice dataset at the outset of the k-mer analysis. Observe that the vast majority of reads now have a signature size in the expected ranges, with the exception of reads that appear in more than 80 pools. This latter set of reads cannot be deconvoluted and is discarded.

The set of reads with a signature of size 7, 12–14 or 15–21 that could be deconvoluted was $\approx 81.5\%$ of the total. Since we knew the BAC from which each read was generated, we determined that 99.57% of the deconvoluted reads were assigned to either the correct BAC or to a BAC overlapping the correct BAC. After deconvolution, the average sequencing depth for each BAC was ≈ 87x, about 50% higher than the expected 56x. Even if we are losing about 18.5% of the reads due to their invalid signatures, deconvoluted reads are frequently assigned to multiple BACs, thereby amplifying the sequencing depth.

In the final step of the protocol, we independently assembled the set of reads assigned to each BAC. We carried out this step with VELVET [13] for each of the 2,197 BACs, for a variety of choices of k-mer size (hash length) and reported only the assembly that maximized the N50[2]. This is an arbitrary choice that does not guarantee the "best" overall assembly. If we average assembly statistics over all the 2,197 BACs, the percentage of reads used in the assembly was 82.3%, the average number of contigs was 41, the average N50 was 47,551 bp (31.4% of the average BAC length), the average largest contig was 57,258 bp (37.8% of the average BAC length), the average sum of all contig sizes was 137,050 bp (90.7% of the average BAC length). The N50 is very high, and so is the percentage of reads used by the assembler. While these numbers already indicate high quality assemblies, we determined whether BACs were correctly assembled by BLAST-ing BAC contigs against the rice genome. Considering these statistics over all the 2,197 BACs, the average BAC coverage was 76.8%, the average gap size was 263 bp, the average number of gaps was 138, the average overlap size was 107 bp, and the average number of overlaps was 75.

To establish a comparison "baseline" for these assembly statistics, we considered the most optimistic scenario of a "perfect deconvolution", which entails using the provenance annotation of each read to assign it back to the correct BAC with 100% accuracy. If we compute the average over all the 2,197 BACs, the

[2]N50 indicates the minimum length of all contig/scaffolds that together account for at least 50% of the genome.

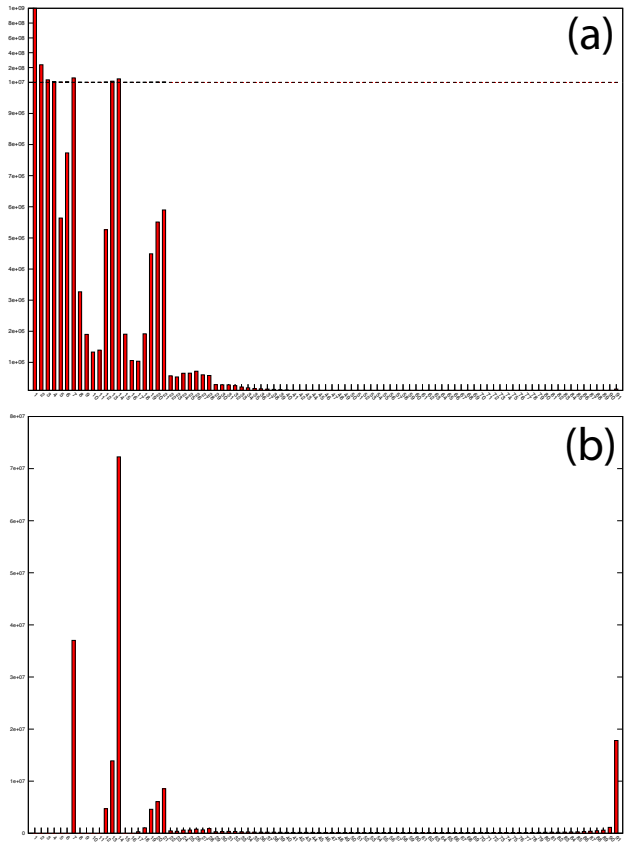

Figure 2: Frequency distribution for the signatures of all distinct 26-mers for the rice synthetic data and all the reads in the 91 pools of sequencing data; the x-axis represents the size of the signature, the y-axis is the frequency.

average fraction of the reads used by VELVET was 82.7% and the average N50 was 132,865 bp (88% of the average BAC length). The BLAST statistics showed an average BAC coverage of 96.3%, an average gap size of 52 bp, an average number of gaps of 97, an average overlap size of 29 bp, and an average number of overlaps of 54. While this latter BAC coverage is about 20% higher, the results following deconvolution compare very favorably with what would be possible sequencing each BAC separately.

3 Educational Achievements and Outreach

In the context of this project, we are currently training one female PhD student (Computer Science), one MS student (Computer Science), and one undergraduate student (Computer Science). Both will be trained in the domain of computational biology.

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.

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 back in his country (Saudi

Arabia).

Undergraduate student Matt Alpert has been working on the BAC-by-BAC assemblies of the rice data. Last summer, I requested and obtained additional funding as an REU to support him. He has applied to several graduate schools, but I hope he will stay at UC Riverside.

During the summer of 2011, post-doc Francesca Cordero (University of Torino, Italy) and post-doc Marco Beccuti (University of Torino, Italy) visited my lab. In collaboration with Francesca, Marco and another faculty in my department (Prof. G. Ciardo) we developed and implemented the deconvolution method described above. Francesca and Marco are back in Italy, but continue to collaborate with us.

Regarding outreach, I have been invited to give a presentation at the workshop "Group Testing Designs, Algorithms, and Applications to Biology" hosted by the Institute for Mathematics and its Applications, University of Minnesota in February 2012. This workshop is by invitation only.

4 Conclusions for Year 1

A summary of the research achievements for Year 1 is as follows.

- We prepared and delivered the libraries for cowpea pools of BACs to the sequencing facility on campus – we should receive data in about a month
- We designed and tested the software tool that deconvolutes the reads to BACs
- Deconvolution results on rice simulated data are very good – 99.57% of the deconvoluted reads are assigned to the correct BAC
- We have set up a pipeline to carry out BAC-by-BAC assemblies using VELVET
- BAC assembly results on the deconvoluted reads have very high quality – BACs are covered by contigs over about 77% of their length, on average

We are confident that the method will be successful on the cowpea data. We have a draft of a manuscript summarizing the method and the simulations. We will write another one on the results on the cowpea genome.

References

[1] BOZDAG, S., CLOSE, T., AND LONARDI, S. A compartmentalized approach to the assembly of physical maps. In *Proceedings of IEEE International Symposium on Bioinformatics & Bioengineering (BIBE'07)* (2007), pp. 218–225.

[2] BOZDAG, S., CLOSE, T. J., AND LONARDI, S. Computing the minimal tiling path from a physical map by integer linear programming. In *Proceedings of the Workshop on Algorithms in Bioinformatics (WABI'08)* (2008), pp. 148–161.

[3] DING, Y., JOHNSON, M. D., CHEN, W. Q., WONG, D., CHEN, Y.-J., BENSON, S. C., LAM, J. Y., KIM, Y.-M., AND SHIZUYA, H. Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics 74*, 2 (2001), 142–154.

[4] ENGLER, F. W., HATFIELD, J., NELSON, W., AND SODERLUND, C. A. Locating sequence on FPC maps and selecting a minimal tiling path. *Genome Research 13*, 9 (2003), 2152–2163.

[5] FERRAGINA, P., AND MANZINI, G. Opportunistic data structures with applications. In *Proceedings of FOCS* (2000), pp. 390–398.

[6] FERRAGINA, P., AND MANZINI, G. An experimental study of an opportunistic index. In *Proceedings of SODA* (2001), pp. 269–278.

[7] LUO, M.-C., THOMAS, C., YOU, F. M., HSIAO, J., OUYANG, S., BUELL, C. R., MALANDRO, M., MCGUIRE, P. E., ANDERSON, O. D., AND DVORAK, J. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics 82*, 3 (2003), 378–389.

[8] NELSON, W. M., BHARTI, A. K., BUTLER, E., WEI, F., FUKS, G., KIM, H., WING, R. A., MESSING, J., AND SODERLUND, C. Whole-genome validation of high-information-content fingerprinting. *Plant Physiology 139*, 1 (2005), 27–38.

[9] SIMPSON, J. T., AND DURBIN, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics 26*, 12 (2010), i367–i373.

[10] THIERRY-MIEG, N. A new pooling strategy for high-throughput screening: the shifted transversal design. *BMC Bioinformatics 7*, 28 (2006).

[11] VÄLIMÄKI, N., LADRA, S., AND MÄKINEN, V. Approximate All-Pairs Suffix/Prefix Overlaps. *Proceedings of Combinatorial Pattern Matching Proceedings of CPM* (2010).

[12] WU, Y., LIU, L., CLOSE, T., AND LONARDI, S. Deconvoluting the BAC-gene relationships using a physical map. In *Proceedings of LSS Computational Systems Bioinformatics Conference (CSB'07)* (August 2007), pp. 203–214.

[13] ZERBINO, D., AND BIRNEY, E. VELVET: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research 8*, 5 (2008), 821–9.