PI: **Batzoglou, SERAFIM**	Title: Read Clouds for Genome Sequencing, Resequencing, and Metagenomics	
Received: 12/17/2008	FOA: PAR09-012	Council: 05/2009
Competition ID: 10	FOA Title: Pre-Application for the 2009 NIH Director's Pioneer Award Program (X02)	
1 X02 OD005175-01	Dual: RM	Accession Number: 3127710
IPF: 8046501	Organization: STANFORD UNIVERSITY	
Former Number:	Department: Computer Science	
IRG/SRG: ZGM1 NDPA-B (01)X	AIDS: N	Expedited: N
Subtotal Direct Costs (excludes consortium F&A)	Animals: N Humans: N Clinical Trial: N Current HS Code: 10 HESC: N	New Investigator: N Early Stage Investigator: N
Senior/Key Personnel:	*Organization:*	*Role Category:*
Serafim Batzoglou	Board of Trustees of the Leland Stanford Junior University	PD/PI

APPLICATION FOR FEDERAL ASSISTANCE

SF 424 (R&R)

2. DATE SUBMITTED	Applicant Identifier
3. DATE RECEIVED BY STATE	State Application Identifier

1. * TYPE OF SUBMISSION

● Pre-application ❍ Application
❍ Changed/Corrected Application

4. Federal Identifier

5. APPLICANT INFORMATION *** Organizational DUNS:**009214214

* Legal Name: Board of Trustees of the Leland Stanford Junior University
Department: Office of Sponsored Research — Division: Engineering
* Street1: 340 Panama Street — Street2:
* City: Stanford — County: Santa Clara — * State: CA: California
Province: — * Country: USA: UNITED STATES — * ZIP / Postal Code: 94305

Person to be contacted on matters involving this application

Prefix:	* First Name:	Middle Name:	* Last Name:	Suffix:
Mr.	Gary		Podesta	

* Phone Number: 650-724-6883 — Fax Number: 650-724-2290 — Email: gplaw@stanford.edu

6. * EMPLOYER IDENTIFICATION NUMBER *(EIN) or (TIN)*:
941156365

7. * TYPE OF APPLICANT
O: Private Institution of Higher Education

8. * TYPE OF APPLICATION: ● New
❍ Resubmission ❍ Renewal ❍ Continuation ❍ Revision

Other (Specify):

Small Business Organization Type
❍ Women Owned ❍ Socially and Economically Disadvantaged

If Revision, mark appropriate box(es).
❍ A. Increase Award ❍ B. Decrease Award ❍ C. Increase Duration
❍ D. Decrease Duration ❍ E. Other *(specify)*:

* Is this application being submitted to other agencies? ❍ Yes ● No
What other Agencies?

9. * NAME OF FEDERAL AGENCY:
National Institutes of Health

10. CATALOG OF FEDERAL DOMESTIC ASSISTANCE NUMBER:
93.310
TITLE: Trans-NIH Research Support

11. * DESCRIPTIVE TITLE OF APPLICANT'S PROJECT:
Read Clouds for Genome Sequencing, Resequencing, and Metagenomics

12. * AREAS AFFECTED BY PROJECT *(cities, counties, states, etc.)*
California

13. PROPOSED PROJECT:

* Start Date	* Ending Date
09/30/2009	07/31/2014

14. CONGRESSIONAL DISTRICTS OF:

a. * Applicant	b. * Project
CA-014	CA-014

15. PROJECT DIRECTOR/PRINCIPAL INVESTIGATOR CONTACT INFORMATION

Prefix:	* First Name:	Middle Name:	* Last Name:	Suffix:
	Serafim		Batzoglou	

Position/Title: Associate Professor — * Organization Name: Board of Trustees of the Leland Stanford Junior University
Department: Computer Science — Division: Artificial Intelligence
* Street1: 353 Serra Mall — Street2: GATES BUILDING 1A-146
* City: Stanford — County: Santa Clara — * State: CA: California
Province: — * Country: USA: UNITED STATES — * ZIP / Postal Code: 94305
* Phone Number: 650-723-3334 — Fax Number: — * Email: serafim@cs.stanford.edu

16. ESTIMATED PROJECT FUNDING		**17. * IS APPLICATION SUBJECT TO REVIEW BY STATE EXECUTIVE ORDER 12372 PROCESS?**
a. * Total Estimated Project Funding	$0.00	a. YES ○ THIS PREAPPLICATION/APPLICATION WAS MADE AVAILABLE TO THE STATE EXECUTIVE ORDER 12372 PROCESS FOR REVIEW ON: DATE:
b. * Total Federal & Non-Federal Funds	$0.00	b. NO ● PROGRAM IS NOT COVERED BY E.O. 12372; OR
c. * Estimated Program Income	$0.00	○ PROGRAM HAS NOT BEEN SELECTED BY STATE FOR REVIEW

18. By signing this application, I certify (1) to the statements contained in the list of certifications* and (2) that the statements herein are true, complete and accurate to the best of my knowledge. I also provide the required assurances * and agree to comply with any resulting terms if I accept an award. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties. (U.S. Code, Title 18, Section 1001)

● * I agree

** The list of certifications and assurances, or an Internet site where you may obtain this list, is contained in the announcement or agency specific instructions.*

19. Authorized Representative

Prefix:	* First Name:	Middle Name:	* Last Name:	Suffix:
	Gary		Podesta	

* Position/Title: Contract and Grant Officer
* Organization Name: Board of Trustees of the Leland Stanford Junior University
Department: Office of Sponsored Research
Division: Engineering
* Street1: 340 Panama Street
Street2:
* City: Stanford
County: Santa Clara
* State: CA: California
Province:
* Country: USA: UNITED STATES
* ZIP / Postal Code: 94305
* Phone Number: 650-724-6883
Fax Number: 650-724-2290
* Email: gplaw@stanford.edu

*** Signature of Authorized Representative**	*** Date Signed**
Gary Podesta	12/17/2008

20. Pre-application File Name: Mime Type:

21. Attach an additional list of Project Congressional Districts if needed.

File Name: Mime Type:

424 R&R and PHS-398 Specific
Table Of Contents

RESEARCH & RELATED Project/Performance Site Location(s)

Project/Performance Site Primary Location

Organization Name: Board of Trustees of the Leland Stanford Junior University

* Street1: 353 Serra Mall Gates		Street2: Building 1A - 146
* City: Stanford	County: Santa Clara County	* State: CA: California
Province:	* Country: USA: UNITED STATES	* Zip / Postal Code: 94305

Project/Performance Site Location 1

Organization Name: Computer Science Department

* Street1: James H. Clark Center		Street2: 318 Campus Drive S266
* City: Stanford	County: Santa Clara County	* State: CA: California
Province:	* Country: USA: UNITED STATES	* Zip / Postal Code: 94305

File Name	Mime Type

Additional Location(s)

RESEARCH & RELATED Other Project Information

1. * Are Human Subjects Involved?	❍ Yes ● No	
1.a. If YES to Human Subjects		
Is the IRB review Pending?	❍ Yes ❍ No	
IRB Approval Date:		
Exemption Number:	_ 1 _ 2 _ 3 _ 4 _ 5 _ 6	
Human Subject Assurance Number		
2. * Are Vertebrate Animals Used?	❍ Yes ● No	
2.a. If YES to Vertebrate Animals		
Is the IACUC review Pending?	❍ Yes ❍ No	
IACUC Approval Date:		
Animal Welfare Assurance Number		
3. * Is proprietary/privileged information included in the application?	❍ Yes ● No	
4.a. * Does this project have an actual or potential impact on the environment?	❍ Yes ● No	
4.b. If yes, please explain:		
4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed?	❍ Yes ❍ No	
4.d. If yes, please explain:		
5.a. * Does this project involve activities outside the U.S. or partnership with International Collaborators?	❍ Yes ● No	
5.b. If yes, identify countries:		
5.c. Optional Explanation:		
6. * Project Summary/Abstract	1851-BAabstract_AS.pdf	Mime Type: application/pdf
7. * Project Narrative	2812-PHRstatement_AS.pdf	Mime Type: application/pdf
8. Bibliography & References Cited		
9. Facilities & Other Resources		
10. Equipment		
11. Other Attachments	7737-SerafimAccomplishments-1.pdf	Mime Type: application/pdf

10 Quantitative and Computational Biology

I propose to develop a novel technology to serve as front-end DNA preparation for virtually any *de novo* sequencing, resequencing, and metagenomics project; the purpose is to enable highly accurate, facile assembly with low computational cost. My method, which I call the read-cloud DNA preparation, is based on the concept of performing hierarchical sequencing without involving an intermediate expensive cloning step. The main idea is that sequenced reads are obtained from long DNA fragments that cover the target genomic region with high redundancy. Each read contains an appended barcode that identifies the source long fragment; reads from the same long fragment form a read cloud. Read clouds cover their source long DNA fragment at low redundancy, and reads from overlapping clouds are pooled together during assembly. Assembly with read clouds is dramatically easier than assembly with whole-genome shotgun mate-pair reads, especially when reads are short. Most sequencing applications will benefit substantially by the read-cloud preparation: de novo sequencing of a complex genome will be possible with short, unpaired reads, while the use of long reads will result in dramatically improved assemblies compared to the traditional shotgun approach. In metagenomics, the read-cloud method will enable for the first time whole-genome microbial assembly of any species that is significantly represented in a metagenomic sample. In human resequencing, my method will allow read mapping at roughly the speed of loading the reads to memory. It will also enable assembly of the significant fraction of the human genome that is currently invisible to short reads because of repeats. Finally, it will enable haplotyping of the sequenced reads. My proposed technology and computational methods are entirely novel. Technology development will be a new direction in my research, and to succeed I will receive extensive mentorship by my colleagues at the Stanford Bioengineering department.

Public Health Relevance Statement

The NIH Pioneer program promotes the development of innovative methods that have the potential to make a major impact on public health. If my proposed DNA preparation method succeeds, it can serve as front-end for virtually any high-throughput DNA sequencing technology and dramatically facilitate assembly and improve the quality of the results. Most importantly, it will have a major impact on human genome resequencing and on cancer genome sequencing. In both these applications, my method will cut computation costs, which currently are becoming the major bottleneck, by a factor of 100 or more; it will enable finding variations in the 10-30% of the human genome that is currently invisible because of repeats; and finally it will allow effective separation of paternal and maternal chromosomes, or haplotypes, which will greatly facilitate studies that associate genetic diseases with their causal genomic variants.

Research Accomplishments

I am most proud of two lines of work in my research: (1) my work on sequence assembly that started when I developed the prototype of the Arachne assembler for my PhD thesis, a system now further developed and maintained at the Broad Institute and which has been used to assemble many complex genomes including the mouse and dog genomes. My recent assembly work includes the development of a proposed sequencing protocol and assembly methodology for de novo assembly of a mammalian genome with short, unpaired reads (see below); this proposal is based on this work. (2) My work on developing discriminative machine learning methods based on conditional random fields for a number of core bioinformatics problems such as gene finding, sequence alignment, and RNA secondary structure prediction (described below). In this line of work I demonstrated how to build systems that are easy to train automatically and rigorously without human tuning, run efficiently, and are significantly more accurate than the previous state of the art. My tools have been widely used by the genomics community.

Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S. Whole-genome sequencing and assembly with high-throughput short-read technologies. *PLOS One,* 2(5): e484. Here we first proposed the read-cloud method for de novo assembly of mammalian genomes with short, unpaired reads. In 2004, witnessing the development of Pyrosequencing and other technologies for high-throughput DNA sequencing, we anticipated the need for methods for de novo assembly of short reads. We proposed that instead of the traditional double-barreled whole-genome shotgun sequencing strategy, the read-cloud method for sequencing should be employed; we suggested that either traditional BAC cloning, or a future technology could be used to produce read clouds. We developed SHRAP (SHort Read Assembly Protocol), an algorithm for de novo assembly of read clouds, and demonstrated through simulations that SHRAP produces draft-quality assemblies of complex eukaryotic genomes such as the human genome. SHRAP is based on first ordering the read clouds and subsequently performing assembly in three stages: (1) local assemblies of regions significantly smaller than the span of a cloud, (2) cloud-sized assemblies of the results of stage 1, and (3) chromosome-sized assemblies. By aggressively localizing the assembly problem during stage 1, our method succeeded in assembling short reads in the presence of repeats. We tested our assembler using simulations on the D. melanogaster and human genomes, and produced almost error-free cloud maps and draft-quality assemblies with reads of length 100-200 bp and 1-3% sequencing errors. In this proposal, I aim to take the theoretical concept in the SHRAP paper, and develop both the enabling technology and broadened computational methods, in order to widely apply read clouds and facilitate assembly for de novo sequencing, resequencing and metagenomics projects.

Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. ISMB 2006 Conference Proceedings, Bioinformatics 22:e90 e98, 2006. *Best Paper Award ISMB 2006.* Together with proteins, RNA structures are the key biomolecules that perform regulatory, catalytic, structural and other biological functions in a cell. RNA structure is determined primarily by the base pairing of nucleotides in the RNA sequence, and computational methods have been among the main ways to predict RNA structure. For several decades, free-energy minimization methods have been the dominant computational strategy. More recently, stochastic context-free grammars (SCFGs) have emerged as an alternative probabilistic method, enabling fully automated statistical learning algorithms to derive model parameters. However, energy minimization methods such as Mfold have clearly dominated probabilistic methods in predicting structures accurately. We developed CONTRAfold, a secondary structure prediction method based on conditional log-linear models (CLLMs), which is the first method enabling automated parameter learning to outperform energy minimization methods, and in particular Mfold which had been the best method for almost two decades. Our result demonstrated that statistical learning procedures provide an effective alternative to laborious experimental measurements of thermodynamic parameters for RNA secondary structure prediction.

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

PROFILE - Project Director/Principal Investigator

Prefix	* First Name	Middle Name	* Last Name	Suffix
	Serafim		Batzoglou	

Position/Title: Associate Professor

Department: Computer Science

Organization Name: Board of Trustees of the Leland Stanford Junior University

Division: Artificial Intelligence

* Street1: 353 Serra Mall

Street2: GATES BUILDING 1A-146

* City: Stanford

County: Santa Clara

* State: CA: California

Province:

* Country: USA: UNITED STATES

* Zip / Postal Code: 94305

*Phone Number	Fax Number	* E-Mail
650-723-3334		serafim@cs.stanford.edu

Credential, e.g., agency login: BATZOGLOU.SERAFIM

*** Project Role:** PD/PI

Other Project Role Category:

	File Name	Mime Type
***Attach Biographical Sketch**	9553-BABio_AS.pdf	application/pdf
Attach Current & Pending Support	4258-BACP_AS.pdf	application/pdf

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

Additional Senior/Key Person Form Attachments

When submitting senior/key persons in excess of 8 individuals, please attach additional senior/key person forms here. Each additional form attached here, will provide you with the ability to identify another 8 individuals, up to a maximum of 4 attachments (32 people).

The means to obtain a supplementary form is provided here on this form, by the button below. In order to extract, fill, and attach each additional form, simply follow these steps:

- Select the "Select to Extract the R&R Additional Senior/Key Person Form" button, which appears below.
- Save the file using a descriptive name, that will help you remember the content of the supplemental form that you are creating. When assigning a name to the file, please remember to give it the extension ".xfd" (for example, "My_Senior_Key.xfd"). If you do not name your file with the ".xfd" extension you will be unable to open it later, using your PureEdge viewer software.
- Using the "Open Form" tool on your PureEdge viewer, open the new form that you have just saved.
- Enter your additional Senior/Key Person information in this supplemental form. It is essentially the same as the Senior/Key person form that you see in the main body of your application.
- When you have completed entering information in the supplemental form, save it and close it.
- Return to this "Additional Senior/Key Person Form Attachments" page.
- Attach the saved supplemental form, that you just filled in, to one of the blocks provided on this "attachments" form.

Important: Please attach additional Senior/Key Person forms, using the blocks below. Please remember that the files you attach must be Senior/Key Person Pure Edge forms, which were previously extracted using the process outlined above. Attaching any other type of file may result in the inability to submit your application to Grants.gov.

1) Please attach Attachment 1	
2) Please attach Attachment 2	
3) Please attach Attachment 3	
4) Please attach Attachment 4	

ADDITIONAL SENIOR/KEY PERSON PROFILE(S)	**Filename** **MimeType**
Additional Biographical Sketch(es) (Senior/Key Person)	**Filename** **MimeType**
Additional Current and Pending Support(s)	**Filename** **MimeType**

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Serafim Batzoglou	**Assistant Professor of Computer Science**
eRA COMMONS USER NAME BATZOGLOU.SERAFIM	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Massachusetts Institute of Technology	B.S.	1996	Mathematics
Massachusetts Institute of Technology	B.S.	1996	Computer Science
Massachusetts Institute of Technology	M.Eng.	1996	E.E.C.S.
Massachusetts Institute of Technology	Ph.D.	2000	Computer Science
Whitehead/MIT Center for Genome Research	Postdoc	2001	Genomics

A. Positions and Honors

Positions and Employment

Apr 2000-Aug 2001 Research Scientist, Whitehead Institute/MIT Center for Genome Research.
Sep 2001-April 2008 Assistant Professor of Computer Science, Department of Computer Science, Stanford University.
May 2008-present Associate Professor of Computer Science, Department of Computer Science, Stanford University.

Other Experience and Professional Memberships

2009 Program Chair, RECOMB Conference
2006 - present Steering Committee, RECOMB Conference
2002, 2004, 2005 Program Committee, RECOMB Conference
2004 - present Editorial Board Member, Genome Research
2002 - 2005 Chair, Program Committee, Annual Satellite RECOMB meeting on DNA Sequencing, Technologies, and Computation
2003 - 2005 Program Committee, IEEE Bioinformatics Conference, CSB2003
2003 Session Chair, Gene Regulation, Pacific Symposium on Biocomputing, PSB2003
2002 Program Committee, Second SIAM International Conference on Data Mining

Honors

2006 Best Paper Award, ISMB 2006 Conference
2004 Best Paper Award, Joint ISMB/ECCB 2004 Conference
2004 Sloan Fellowship, Molecular and Computational Biology
2004 NSF CAREER Award, “Methods for Comparative Genomics”
2003 Top 100 Technology Innovators Award, Technology Review magazine, MIT
2002 James H. Clark Faculty Scholar, School of Engineering, Stanford University
1998 - 2000 Merck/MIT Graduate Fellowship
1997 - 1998 Program of Mathematics and Molecular Biology Graduate Fellowship

B. Selected Peer-reviewed publications (in chronological order) (from a total of 30 peer-reviewed journal articles, 22 peer-reviewed conference articles, and 5 chapters and solicited articles).

Batzoglou S, Mesirov JP, Berger B, Lander ES. Sequencing a Genome by Walking with Clone-ends: A Mathematical Analysis. Genome Research 9:1163-1174, 1999.

Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. Genome Research 10:950-958, 2000.

Batzoglou S, Istrail S. Physical Mapping with Repeated Probes: The Hypergraph Superstring Problem. Journal of Discrete Algorithms 1(1):51-76, 2000.

International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome. Nature 409:860-921, 2001.

Batzoglou S, Jaffe D, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. ARACHNE: A Whole Genome Shotgun Assembler. Genome Res. 12:177-189, 2002.

Brudno M, Do C, Cooper GM, Kim MF, Davydov E, NISC Comparative Sequencing Program, Green ED, Sidow A, Batzoglou S. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13:721-731, 2003.

Cooper GM, Brudno M, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Res. 13:813-820

Khambata Ford S, Liu Y, Gleason C, Dickson M, Altman RB, Batzoglou S, Myers RM. Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. *Genome Research* **13:**1765-1774, 2003.

Brudno M, Malde S, Poliakov A, Do C, Couronne O, Dubchak I, Batzoglou S. Glocal alignment: finding rearrangements during alignment. *Special Issue on the Proceedings of the ISMB 2003*, *Bioinformatics* **19:** 54i-62i, 2003.

Lee S-I, Batzoglou S. Application of independent component analysis to microarrays. *Genome Biology* **4:**R76, 2003.

Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S. Eukaryotic regulatory element conservation and their identification using comparative genomics. *Genome Research* 14:451-458, 2004.

Cooper GM, Brudno M, Stone ES, Dubchak I, Batzoglou S, Sidow A. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Research* 14:539–548, 2004.

Rat Genome Sequencing Project Consortium (RGSPC). Genome sequence of the Brown Norway Rat yields insights into mammalian evolution. *Nature* 428:493–521, 2004.

Brudno M, Poliakov A, Salamov A, Cooper GM, Sidow A, Rubin EM, Solovyev V, Batzoglou S, Dubchak I. Automated whole-genome multiple alignment of Rat, Mouse, and Human. *Genome Research* 14:685–692, 2004.

Do CB, Brudno M, Batzoglou S. ProbCons: probabilistic consistency-based multiple alignment of amino acid sequences. Proceedings of the 12th International Conference on Intelligent Systems for Molecular Biology & 3rd European Conference in Computational Biology, 2004. **Best Paper Award.** Full paper, Genome Research 15:330–340 , 2005.

The ENCODE Project Consortium. The ENCODE Project. *Science* 306:636–640, 2004.

Batzoglou S. The many faces of sequence alignment. Briefings in Bioinformatics 1: 6–22, 2005.

Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* 15: 901–913, 2005.

Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman J, Batzoglou S, et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature,* 438: 1105–1115, 2005.

Fratkin E, Naughton B, Brutlag DL, Batzoglou S. MotifCut: Finding Regulatory Motifs with Maximum Density Subgraphs. *Special Issue on the Proceedings of the ISMB2006*, *Bioinformatics* 22: e150-e157, 2006.

Do CB, Gross SS, Batzoglou S. CONTRAlign: Discriminative Training for Protein Sequence Alignment. *Proceedings of the Tenth Annual RECOMB Conference, 2006, pp.* 160–164.

Srinivasan B, Novak A, Flannick J, Batzoglou S, McAdams H. Integrated Protein Interaction Networks for 11 Microbes. *Proceedings of the Tenth Annual RECOMB Conference, 2006, pp.* 1–14.

Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S. Graemlin: general and robust alignment of multiple large interaction networks. *Genome Research,* 16:1169-1181, 2006.

Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA Secondary Structure Prediction without Physics-Based Models. *Special Issue on the Proceedings of the ISMB2006*, *Bioinformatics* 22:e90-98, 2006*.* **Best Paper Award.**

Margulies E, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Research,* 17(6): 760-774, 2007.

The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816, 2007.

Drosophila Comparative Genome Sequencing and Analysis Consortium. Evolution of genes and genomes in the context of the Drosophila phylogeny. *Nature,* 450:203-218, 2007.

Do CB, Batzoglou S. What is the EM algorithm? *Nature Biotechnology, 26:897-899, 2008.*

Sundquist A, Fratkin E, Do CB, Batzoglou S. Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Research 18:676-682,2008.

Deshpande O, Batzoglou S, Feldman M, Cavalli-Sforza L. A serial founder effect model for human settlements out of Africa. *Proceedings of the Royal Society B,* 276:291-300, 2008.

Valouev A, Johnson DS, Sundquist A, Medina C, Elisabeth A, Batzoglou S, Myers RM, Sidow A. Genome-wise analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods, in press*

Current and Pending Support - Serafim Batzoglou

ACTIVE

DBI-0347952-004 (Batzoglou) 03/15/04-02/28/09 0.75 mo Summer
NSF- National Science Foundation: CAREER $557,132.00
Methods for Comparative Genomics

Aims: Develop algorithms, systems, and statistical analysis methods for comparative genomics. The core bioinformatics problems in which this proposal focuses are DNA and protein sequence multiple alignment, gene recognition based on comparison of the human genome and other mammalian genomes, and statistical methods for modeling the evolution of genomic sequences.

DBI- 0640211-001(Batzoglou) 08/15/07-07/31/10 0.75 mo Summer
NSF- National Science Foundation $998,708.00
Protein Interaction Networks: Interaction Alignment

Aims: Recently, various high-throughput experimental assays as well as computational methods have been developed that provide us information on which groups of proteins work together in modules or pathways within a cell. Networks of protein associations summarize such information by representing proteins as nodes, and interactions, or associations between pairs of proteins as edges. In this project undertake a large-scale effort to integrate experimental and computational data into protein association networks for the majority of currently sequenced microbial species. We then develop computational methods for comparing these networks across species, and develop a robust and efficient tool for protein network multiple alignment.

R01-HG003685-4 (Batzoglou) 09/01/05-08/31/09 2.00 mo Cal
NIH- National Institutes of Health $1,508,818
Annotation of Constrained Elements in the Human Genome

Aims: In this project, we will first develop computational methods and a high-throughput system for whole-genome mammalian alignment, and apply these methods to the alignment of the human genome to all mammalian genomes that are available in both high-quality and draft form. Based on these alignments, the goal is to find elements on the human genome that are constrained by evolution to be conserved among all mammalian species. We will also address constraint at the level of vertebrate by including chicken and several fish genomes in our analysis.

7U54 HG004576-02
NIH- National Institutes of Health-
Hudson-Alpha Institute for Biotechnology 07/1/08 -6/30/09 0.45 mo aca. .15 Sum
Encode - Regulatory Networks $213,416

Aims: The role of Dr. Batzoglou in the ENCODE 2 project is to apply cross-species comparison methods and identify sequence elements on the human genome that are conserved across multiple species, and to integrate multiple gene expression and epigenetic data sources produced by the Stanford group and other groups into models of gene regulation.

PENDING

NSF- National Science Fundation 01/01/09-12/31/11 .90 mo Aca /1 mo sum.
Annotation of Transcripts in Multiple Vertebrate Genomes

Aims: In this project we propose to develop and apply on a large scale methods for annotating protein-coding gene and functional noncoding RNA gene transcripts on vertebrate genomes. Our methods are based on robust and efficient machine learning methodology that integrates comparative sequence information across multiple vertebrate genomes into unified discriminative models of gene structure. We propose to annotate the human genome as well as all other sequenced vertebrates, and to verify experimentally a random subset of our predictions in a set of model organisms.

Statement of Dedication of Effort:

Dr. Batzoglou pledges to devote more than 51% of his total effort to Pioneer program-funded research, if an award is made. Currently Dr. Batzoglou is funded at less than 50% in all his current and pending awards for the 12-month calendar year.

Dr. Batzoglou laboratory has office space for 14 students or Post Docs. He has access to powerful computational resources including the 2,200-core Dell BioX Cluster, funded by the NSF and shared among BioX-affiliated laboratories. Dr Batzoglou is part of the Artificial Intelligence Laboratory of the Computer Science Department, where has access to administrative support and to systems administration for computing.

PHS 398 Cover Page Supplement

OMB Number: 0925-0001
Expiration Date: 9/30/2007

1. Project Director / Principal Investigator (PD/PI)

Prefix:
* First Name: Serafim
Middle Name:
* Last Name: Batzoglou
Suffix:

* New Investigator? ● No ○ Yes

Degrees: PhD

2. Human Subjects

Clinical Trial? ● No ○ Yes

* Agency-Defined Phase III Clinical Trial? ○ No ○ Yes

3. Applicant Organization Contact

Person to be contacted on matters involving this application

Prefix: Mr.
* First Name: Gary
Middle Name:
* Last Name: Podesta
Suffix:

* Phone Number: 650-724-6883
Fax Number: 650-724-2290
Email: gplaw@stanford.edu

* Title: Contract and Grant Officer

* Street1: 340 Panama Street
Street2:
* City: Stanford
County: Santa Clara County
* State: CA: California
Province:
* Country: USA:
* Zip / Postal Code: 94305

PHS 398 Cover Page Supplement

OMB Number: 0925-0001
Expiration Date: 9/30/2007

4. Human Embryonic Stem Cells

* Does the proposed project involve human embryonic stem cells? ●No ❍Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://stemcells.nih.gov/registry/index.asp . Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

Cell Line(s): ❑ Specific stem cell line cannot be referenced at this time. One from the registry will be used.

OMB Number: 0925-0001
Expiration Date: 9/30/2007

PHS 398 Research Plan

1. Application Type:

From SF 424 (R&R) Cover Page and PHS398 Checklist. The responses provided on these pages, regarding the type of application being submitted, are repeated for your reference, as you attach the appropriate sections of the research plan.

*Type of Application:

● New ❍ Resubmission ❍ Renewal ❍ Continuation ❍ Revision

2. Research Plan Attachments:

Please attach applicable sections of the research plan, below.

1. Introduction to Application (for RESUBMISSION or REVISION only)	
2. Specific Aims	
3. Background and Significance	
4. Preliminary Studies / Progress Report	
5. Research Design and Methods	7859-BAEssay_AS.pdf
6. Inclusion Enrollment Report	
7. Progress Report Publication List	

Human Subjects Sections

Attachments 8-11 apply only when you have answered "yes" to the question "are human subjects involved" on the R&R Other Project Information Form. In this case, attachments 8-11 may be required, and you are encouraged to consult the Application guide instructions and/or the specific Funding Opportunity Announcement to determine which sections must be submitted with this application.

8. Protection of Human Subjects	
9. Inclusion of Women and Minorities	
10. Targeted/Planned Enrollment Table	
11. Inclusion of Children	

Other Research Plan Sections

12. Vertebrate Animals	
13. Select Agent Research	
14. Multiple PI Leadership	
15. Consortium/Contractual Arrangements	
16. Letters of Support	
17. Resource Sharing Plan(s)	

18. Appendix

Attachments

IntroductionToApplication_attDataGroup0

File Name	Mime Type

SpecificAims_attDataGroup0

File Name	Mime Type

BackgroundSignificance_attDataGroup0

File Name	Mime Type

ProgressReport_attDataGroup0

File Name	Mime Type

ResearchDesignMethods_attDataGroup0

File Name	Mime Type
7859-BAEssay_AS.pdf	application/pdf

InclusionEnrollmentReport_attDataGroup0

File Name	Mime Type

ProgressReportPublicationList_attDataGroup0

File Name	Mime Type

ProtectionOfHumanSubjects_attDataGroup0

File Name	Mime Type

InclusionOfWomenAndMinorities_attDataGroup0

File Name	Mime Type

TargetedPlannedEnrollmentTable_attDataGroup0

File Name	Mime Type

InclusionOfChildren_attDataGroup0

File Name	Mime Type

VertebrateAnimals_attDataGroup0

File Name	Mime Type

SelectAgentResearch_attDataGroup0

File Name	Mime Type

MultiplePILeadershipPlan_attDataGroup0

File Name	Mime Type

ConsortiumContractualArrangements_attDataGroup0

File Name	Mime Type

LettersOfSupport_attDataGroup0

File Name	Mime Type

ResourceSharingPlans_attDataGroup0

File Name	Mime Type

Appendix

File Name	Mime Type

Read Clouds for Genome Sequencing, Resequencing, and Metagenomics

PROJECT DESCRIPTION.

DNA sequencing is undergoing a technological revolution that dramatically increases throughput and decreases cost per base. The new sequencing technologies that are already in wide use, such as the 454, Solexa, and Solid platforms, deliver several Gbp of sequence per week in the form of short fragments, or *reads*; next generation technologies promise even higher throughputs. As a result of this dramatic cost reduction, many applications are enabled that were impractical until now: personal genome sequencing, sequencing of metagenomic samples, and *de novo* sequencing of species that are not model organisms. However, the sequenced reads need to be assembled computationally in order to be of use. Because of the limits of the new technologies in read length and read quality, and because sequencing throughput increases much faster than Moore's law, the computational assembly problem is becoming increasingly critical: cost of computational infrastructure and personnel is quickly dominating over reagent cost, while assembled genomes are far from perfect – for example, in personal genome sequencing, a large part of the human genome is invisible due to repeats.

I propose to develop a new technology to serve as front-end DNA preparation for virtually any *de novo* sequencing, resequencing, and metagenomics project. My method, which I call the *read-cloud* DNA preparation, is designed to enable dramatically easier sequence assembly; it will reduce the computational cost of assembly by orders of magnitude in some applications, while making the assembled genomes significantly more complete and accurate. The read-cloud method is based on the concept of performing hierarchical sequencing without involving an intermediate expensive cloning step. Most sequencing applications will benefit substantially by my method: ***de novo* sequencing** of a complex genome will be possible with short, unpaired reads, while the use of long reads will result in dramatically improved assemblies compared to the traditional double-barreled shotgun approach. The impact in **metagenomics** will be even greater: any microbe that is significantly represented in a sample containing the genomes of multiple species will be assembled, whereas today whole-genome assembly of microbes in a metagenomics samples is virtually impossible. Finally, **human resequencing** becomes much easier: first, the computational mapping problem will require a few orders of magnitude less CPU time; second, most of the repetitive portion of the human genome that today cannot be mapped with short reads will be easy to reconstruct; third, haplotyping will become straightforward.

I first described the basic computational protocol in a paper focusing on whole-genome *de novo* mammalian assembly with short, unpaired reads [3]. In that paper, I proposed that BAC cloning or some other future technology might be used to implement my protocol. Then, I very recently described this concept to my colleague from Bioengineering, Annelise Barron over coffee at Peets in Stanford's James H. Clark Center in which we both work and have our labs. She immediately found the idea extremely exciting and told me that my sequencing protocol can become a reality technologically, if new approaches to preparing and handling DNA are used. Since then, I have talked about the idea in detail with my colleagues Steve Quake and Ron Davis, and had many extensive brainstorming discussions with Annelise Barron. All of my colleagues, and Annelise especially, are excited to help me make this technology a reality, while in parallel I will develop new assembly methods that will utilize the resulting sequencing data. This would be a first: I will be working to develop a new DNA handling technology exclusively *for the purpose of enabling powerful assembly algorithms.*

The read-cloud method for DNA sequencing. The main idea of the read-cloud method is that reads are obtained from long DNA fragments that cover the target genomic regions; each read contains an appended barcode that identifies the source read cloud (Figure 1). Long fragments cover the target genomes with high redundancy so as to avoid gaps; read clouds cover the long fragments lightly because reads from multiple overlapping clouds are aggregated computationally during assembly. DNA sequencing is thus performed in a hierarchical fashion that dramatically facilitates the subsequent assembly and mapping of the reads. Read clouds need not, but may be combined with mate paired reads to further facilitate assembly; reads should be long enough to fit the barcode (~12 bp) plus an informative segment from the genome (>20bp); a read length of 35 bp or longer should suffice.

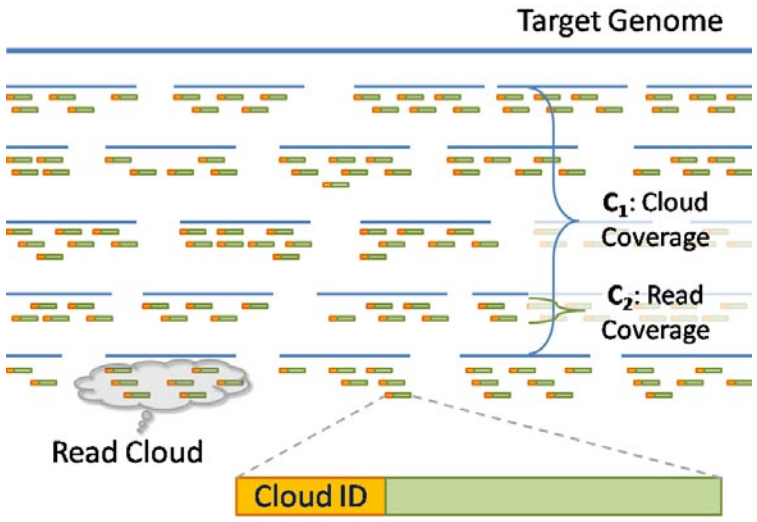

Figure 1. The Read Cloud Method for DNA Preparation. A collection of large DNA fragments (blue) cover at depth C_1 the target genome. From each large fragment, a cloud of reads marked by the unique large fragment ID, and covering the fragment at approximate depth C_2, is obtained. C_1 has to be high (>7x) so as to cover the target genome fully; C_2 may be low (<2x) because reads from multiple overlapping clouds are aggregated computationally during assembly.

***De novo* sequencing and assembly.** The main challenge in the *de novo* sequencing and assembly of a complex eukaryotic genome is the presence of repeats within the genome. Repeats of various lengths (representing ~50% of the human genome), now confound assembly because string overlaps between pairs of reads do not unambiguously imply that the reads come from overlapping locations of the source genome; instead, the reads may cover different copies of the same repeat, in which case merging them could result in misassembled sequence that is not present in the source genome. Sequencing with mate pairs (i.e., with pairs of reads that originate from the two ends of a larger piece of DNA) provides additional information to jump across repeats. Today, draft assemblies of mammalian genomes can be obtained with reasonable quality by performing 7x sequencing with Sanger-length paired reads (~650 bp), as long as mate pairs are obtained from appropriate insert lengths so as to give long-range information to reconstruct the genome across repeats. However, the *de novo* assembly of a complex genome is still considered impossible with unpaired reads, or with reads that are substantially shorter than those provided by Sanger sequencing. Therefore the new, ultra-high throughput technologies are now limited in their applicability.

As we have shown with computational simulations [3], a mammalian genome can be assembled from read clouds with relative ease and high accuracy, so as to obtain draft-quality assemblies, even with unpaired reads of length 100-200 bp and 1-3% sequencing error rate. In our simulations, we covered the human genome at 7-10x depth with long fragments of length 100-200 kb, and covered each fragment with a read cloud of depth 1.5-2x. Although we have not yet performed simulations with paired reads, I anticipate that their combination with read clouds will improve assembly even further.

Briefly, the following assembly algorithm can be used for *de novo* assembly with read clouds: (1) Read clouds are mapped with respect to each other. The main idea in this mapping step is to

maximize the number of unique k-mers that are present in the overlaps between consecutive read clouds. This step results in virtually perfect maps of read clouds, with no errors in our simulated Drosophila melanogaster map, and 98-238 errors in the map of the entire human genome [3]. (2) Reads are collected in bins; each bin consists of a region defined by the boundaries between consecutive read clouds, and is on average 20 kb in length. Then, reads within each bin individually are assembled with a high-accuracy assembler like Euler [1]. (3) For each read cloud, the contigs from step 2 in all bins overlapping this cloud are collected and assembled with Euler to obtain longer contigs. (4) The resulting contigs are assembled according to the map obtained in Step 1. This entire procedure is trivially parallelizable. Using short, unpaired reads it provides draft-quality assemblies of the human genome; if longer reads, or paired reads are used, I expect this procedure to result in assemblies that are of substantially higher accuracy than the draft assemblies now produced by whole-genome shotgun sequencing projects, which will represent a significant advance.

Metagenomics. One of the main purposes of metagenomics projects is to sequence organisms that are difficult to isolate and culture. Today, metagenomics samples that are sequenced with the shotgun method can be assembled in at most operon-size contigs, if at all. Whole microbial genome assembly is impossible when several genomes are present in the source sample, because of repeats, shared operons, and difficulty to separate close strains. By applying the read-cloud DNA preparation, I anticipate that the vast majority of microbes that are nontrivially present in a sample will be assembled into high-quality whole genomes that can be easily distinguished from one another.

The following algorithm can be used for assembly: (1) Cluster the read clouds into groups that come from the same genome. The signature of two clouds coming from overlapping regions of the same microbial genome is that a large fraction of their reads will overlap almost perfectly, within the errors of the technology used for sequencing. In contrast, clouds coming from similar regions of distinct microbes will contain a large fraction of reads whose overlaps contain base-pair differences in the sites of polymorphism. Separation of strains will be possible whenever the polymorphism rate across strains is statistically distinguishable from the sequencing error rate. (2) Assemble each group individually using the same read-cloud assembly algorithm as in *de novo* sequencing. How much sequencing do we need to do, in order to assemble microbes present at low frequency in a sample? Assume that we perform 30 Gbp of sequencing, or the equivalent of 10x of the human genome. Assuming a microbial genome size of 5 Mbp, this provides 6,000x coverage of the sample, or 6x coverage for microbes present at 0.1% frequency in the sample; such microbes will therefore be assembled.

Human Genome Resequencing. Current high-throughput sequencing technologies for the resequencing of human individuals present us with several key challenges: (1) Mapping short reads (e.g., 35-50 bases for Illumina, Solid, and 230-400 bases for 454) onto a reference human genome requires significant computational resources, of roughly 1 CPU day per 1 Gbp of sequenced reads. (2) Depending on the read length, and on whether reads are paired, a large fraction of the human genome is invisible because it is composed of high-fidelity repeat copies, segmental duplications, or copy number variations. (3) Phasing and haplotyping is virtually impossible to do correctly if at all. A successful implementation of the read-cloud method will address all three challenges effectively.

The following algorithm can be used to map read clouds: (1) Each read cloud is rapidly and accurately mapped to its source location in the human genome. As we demonstrated [3] there is enough information within the k-mers of read clouds, so that all but roughly 100-200 read clouds will be

correctly mapped to the entire human genome. The computation for this step is simply to obtain a list of k-mers within the reads of the read cloud, and then count up their matches to each location of the genome (can be done with a range query in linear time in practice, so this step is trivial). (2) Reads are then mapped to their source location, within the restricted ~200 kb region where the cloud maps to the reference genome. This step is also rapid, and enables mapping of reads to most repeats; recent duplications longer than 200 kb, low complexity DNA, tandem repeat arrays, and repeats occurring with high sequence identity within one 200 kb region will still be hard to map. (3) Variations with respect to the reference are detected. (4) Variations are phased, and haplotypes are obtained, because each read cloud comes from one haploid genome. Because there are around 200 or more SNPs and other variants within a 200 kb region, this step is also straightforward.

Technology development. Starting with the target genome or metagenomic sample, the following novel procedure prepares DNA for sequencing so that each of the resulting reads will start with a unique barcode indicating its source cloud.

Step 1. Create large, roughly 200 kb DNA molecules by a focused shear microfluidic degradation of the target genome—for example a human genome or a metagenomic sample.

Step 2. Collapse these large DNA molecules into tight random-coil “balls” with cationic salts, dilute them appropriately, and within a microfluidic device, entrain each DNA nanoball within a water droplet that is formed within a flowing mineral oil phase [2,4]. These droplets will be formed when three streams, two oil streams and a central aqueous stream merge within the device; this technique has been demonstrated by Steve Quake's laboratory and by many others, and is the basis for technology being developed by the company Raindance Technologies. The droplet encapsulation of condensed DNA molecules will ensure that they remain separated from each other and protected from stretching in hydrodynamic flow, which might otherwise break them.

Step 3. Sort the condensed molecules hydrodynamically into separate micro- or nanowells, so that one large DNA molecule is present per well.

Step 4. Apply a random amplification step to each large DNA fragment separately in each well. This process is similar to the one now being used by Steve Quake to amplify single whole bacterial genomes that are isolated with microfluidics, and with care, amplification bias can be minimized (Steve Quake, personal communication).

Step 5. With each DNA family being handled distinctly, again apply focused hydrodynamic shearing to break the amplified DNA molecules down to the size of maximum usefulness for the sequencing technology that will be used subsequently (e.g., 450-500 bases for 454; <100 bases for Illumina).

Step 6. Apply blunt-end ligation to the resulting sheared DNA pieces, to append the following sub-sequences: (1) a 12 bp barcode sequence that is designed algorithmically with redundancy and labels all of the progeny fragments as originating from its specific read cloud; (2) sites for cyclization and a universal site for priming a rolling-circle amplification. Because at most 10^6 distinct clouds will be needed in the foreseeable future, a 10 bp barcode would be sufficient if we assumed no read errors. However with 12 bp barcodes, we will computationally build checksum error correction so that the majority (15/16) of reads that contain a sequencing error within their barcode will be rejected from the initial mapping and assembly phases, and only added optionally later. Such erroneous reads are essentially whole-genome shotgun reads because their source cloud is unknown.

Step 7. All DNA fragments, so far isolated in one well per cloud, are now pooled together. For sequencing technologies like Helicos, Solexa, or 454, at this point the DNA is passed to their

preparation protocol. Optionally, to perform Sanger sequencing, the DNA fragments are amplified once more with rolling-circle amplification.

Up until now, I have worked as a computational scientist only. However, soon I will be courtesy faculty in Bioengineering, and I can recruit BioE students with strong backgrounds in experimental genomics. Annelise Barron and Steve Quake both offer me the use of their wet laboratories and their consultancy about technological details for the use of my Ph.D. students who will work to make these ideas a reality.

EVIDENCE OF INNOVATIVENESS. My research has been innovative in a range of problems in computational genomics. Highlights include: (1) development of the prototype of the Arachne whole-genome shotgun assembler, which today is one of the best large-scale assemblers in use; (2) development of LAGAN, the first efficient large-scale multiple genome aligner, which has above 400 references and is still widely used; (3) introduction of the conditional training methodology in genomics and development of CONTRAST, currently the best *de novo* gene finder, CONTRAfold, the first RNA folding program to outperform energy minimization methods and currently most accurate method, and ProbCons/CONTRAlign, some of the most accurate protein alignment programs to date; (4) development of Graemlin, the first true multiple protein interaction network aligner, which provided orders-of-magnitude speedup and improved accuracy over previous methods; (5) first conception of the read-cloud method a few years ago and development of an algorithm for *de novo* assembly of a mammalian genome with short reads. The proposed novel technology here is a simple concept that will dramatically facilitate a wide range of core DNA sequencing applications.

HOW THE PLANNED RESEARCH DIFFERS FROM MY PAST OR CURRENT WORK. So far my research has focused on developing computational methods for genomics. DNA sequence assembly has been one of my areas of greatest expertise. My proposed research sets a new direction because I will for the first time be developing new technology for the purpose of enabling powerful sequence assembly methods.

SUITABILITY FOR THE PIONEER AWARD PROGRAM. The Pioneer program promotes the development of innovative methods that have the potential to make a major impact on human health. If my proposed method succeeds, it will be applicable as a front end DNA preparation method for the vast majority of sequencing technologies and dramatically facilitate key sequencing applications. In human genome resequencing, my method will enable haplotyping, reduce computation cost of mapping by two to three orders of magnitude, and enable variation detection in a large fraction of the human genome that is currently invisible. In metagenomics, my method will be the first to enable whole-genome assembly of microbes from a mixed sample. Both of these application areas have potentially enormous impact in human health. This project is extremely challenging, the required technological development is in an area where I have little expertise - even though I will receive mentorship from some of the best people in the world on this area - and therefore the project has a significant chance of failing. Therefore, it is unsuitable for a regular NIH R01 and the Pioneer program is the right place for this project.

(1) Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. PNAS 17:9748-9753, 2001. **(2)** Squires TM, Quake SR. Microfluidics: fluid physics at the nanoliter scale. Rev. Mod. Phys. 77: 977, 2005. **(3)** Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S. Whole-genome sequencing and assembly with high-throughput short-read technologies. PLOS One, 2(5): e484, 2007. **(4)** Tice JD, Song H, Lyon AD, Ismagilov RF. Formation of droplets and mixing in multiphase microfluidics at low values of the Reynolds and the capillary numbers. Langmuir 19:9127-9133, 2007.