UNIVERSITY OF CALIFORNIA, BERKELEY

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO   SANTA BARBARA • SANTA CRUZ

LIOR PACHTER
PROFESSOR
MATHEMATICS, MOLECULAR & CELL BIOLOGY
AND COMPUTER SCIENCE
UC BERKELEY

970 EVANS HALL
BERKELEY, CALIFORNIA 94720-3840
TELEPHONE (510) 642-2028
FAX (510) 642-8204
E-MAIL lpachter@math.berkeley.edu
WEBSITE http://math.berkeley.edu/~lpachter

September 3, 2013

Dear Barbara,

I'm writing to follow up on our phone call on Friday (August 30th) to summarize my concerns about proposal 2R01HG006102-4 (renewal of HG006102) . As I explained to you, the description of the PIs contribution to the Cufflinks project (progress report for Aim 3), and the fact that I am not mentioned in the proposal (despite leading the Cufflinks project for 5 years now) shocked me. In addition to that, many claims in the proposal are dishonest and/or false:

1. The PIs repeatedly take credit for prior work that is not theirs.
2. In a progress report describing previous work on NIH R01 HG006102, a paper is listed that neither of the PIs on the grant, or their funding, had anything to do with.
3. Specific Aim 3 includes a project that was already funded in one of my current active grants (a proposal that one of the PIs is fully aware of). The work has already been done.

I elaborate on these issues in this letter, but first provide some background on the Tuxedo Suite (the tools Bowtie, TopHat and Cufflinks) and explain how the Cufflinks project was started and developments in the project over the past 3 years.

History of the Cufflinks Project

In October 2008, Ben Langmead and Cole Trapnell (students of Steven Salzberg at the University of Maryland) submitted the paper on Bowtie, a tool for fast short read mapping to a genome. Although I had nothing to do with the Bowtie project, I had started discussing short read mapping with Cole Trapnell a year earlier (Fall 2007) when he visited me at UC Berkeley to discuss the possibility of visiting for a year. The reason for his visit was so that he both collaborate with me, and also be co-located with his girlfriend at the time (now his wife) who was a Ph.D. student in mathematics in my department. In 2007 we had agreed to think about extensions to Bowtie for spliced read alignment, so that we could map reads from RNA-Seq, a technology that was just coming online.

Cole arrived in Berkeley in September 2008. He had started thinking about spliced read alignment during the previous year after our initial conversation and had written some code, but had also been busy with the Bowtie project. We began working intensely on the project immediately after his arrival, and by late October had submitted a manuscript for publication (published in 2009):

- C. Trapnell, L. Pachter and S. Salzberg, TopHat: discovering splice junctions with RNA-seq, Bioinformatics 25, (2009) 1105–1111.

We agreed that Steven would be last author for two reasons: he had contributed to the project during some conversations with Cole early in 2008, and he was also paying Cole's salary during his year-long visit to Berkeley. Notably however, Cole Trapnell was designated corresponding author by us, an unusual step we took because of his leading role not only in writing code, but developing the algorithm. We named the program TopHat consistently with the "Tuxedo" theme because TopHat relied heavily on the Bowtie tool.

Shortly after TopHat was submitted, Cole and I started to discuss an expanded RNA-Seq project, including the development of an assembler for RNA-Seq based on the TopHat alignments, possibly with a method for quantification of expression. We worked hard on this for the remainder of 2008 and early in 2009, so that by the summer of 2009 we already had a functioning tool for RNA-Seq. At that point, my collaboration with Cole was going so well that he asked his advisor Steven Salzberg whether he could remain in Berkeley for the remainder of his Ph.D. with both of us has his coadvisors. Steven agreed, and after obtaining permission from a dean at U. Maryland (with an explanation that I would officially co-advise Cole), by September 2009 we had submitted a paper on RNA-Seq to Nature Biotechnology. Cole and I collaborated closely with Ali Mortazavi (then at Caltech, now at UC Irvine) and Barbara Wold who generously provided us with RNA-Seq data, and the final paper described our tool Cufflinks:

- C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold and L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nature Biotechnology 28, (2010), p 511–515.

During the Cufflinks project Steven was a minor participant, having conversed with Cole about the project only a few times in person during the year, and occasionally during phone calls. However he was still paying Cole's salary, and we all agreed he should be a coauthor on the paper. I believe his role was accurately represented as third author from last on the paper.

Cole received his Ph.D. in the summer of 2010 with me and Steven Salzberg as official co-advisors (this designation for me was nontrivial and required a formal process at the U. Maryland). Cole left my group and went to work as a postdoc in John Rinn's lab at Harvard University. In the meantime I had submitted a grant to continue development of Cufflinks, and Cole and I agreed to continue our collaboration. At the same time, Steven Salzberg and Rafael Irrizary submitted a grant for further development of Bowtie and TopHat (HG006102). In terms of Cufflinks, we all agreed that Cole and I would continue development of the program, and that Steven's group would provide support in the form of continuing to host the website, providing a programmer to fix bugs, and to support key Cufflinks-related scripts, particularly Cuffcompare (attached at the end of this letter is Steven Salzberg's support letter for my R01 making these points). In addition we confimed them in a phone call with program officer Peter Good before our awards were disbursed. The arrangement worked well and Cole and I continued our fruitful collaboration between Boston and Berkeley. In fact we continue to work together to this day.

After Cole's departure we successfully coauthored *three* more "Cufflinks publications", i.e. papers describing improvements to the original Cufflinks software program. These papers are:

- A. Roberts, C. Trapnell, J. Donaghey, J.L. Rinn and L. Pachter, Improving RNA-Seq expression estimates by correcting for fragment bias, Genome Biology, 12 (2011), R22.
- A. Roberts, H. Pimentel, C. Trapnell and L. Pachter, Identification of novel transcripts in annotated genomes using RNA-Seq, Bioinformatics, 27 (2011), p 2325–2329.
- C. Trapnell, D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn and L. Pachter, Differential analysis of gene regulation at transcript resolution with RNA-seq, Nature Biotechnology, 31 2012), p 46–53.

Neither Steven Salzberg nor anyone from his group have been a coauthor on any of these papers. That is because they simply did not work on any of the projects. Adam Roberts, first author on two of the papers, is my student (about to graduate), and became intimately involved with the Cufflinks project, both making substantial fundamental improvements and changes to the code, and introducing algorithmic developments as described in the papers above. The third project, describing our new tool for differential expression (Cuffdiff2) was a close collaboration with the Rinn lab who produced an extraordinary dataset for testing the statistical methods in the paper. John Rinn is properly credited as co-corresponding author with me on that paper.

The HG006102 renewal

To elaborate on the points made in the beginning of this letter, I draw attention to statements made in the HG006102 proposal (quotes are in italics below). In considering these statements, it is important to note that neither I, nor anyone in my group, nor John Rinn or anyone in his group, are mentioned even a single time in the proposal. Only Cole Trapnell is mentioned once by name (parenthetically). Thus, "we" refers to Steven Salzberg and previous co-PI Rafael Irrizary and/or current co-PI Ben Langmead.

For example, we added a major new module to Cufflinks that allows users to provide annotation, and that uses this annotation as a guide to assembly (the original system assembled transcripts without regard to known genes).

This refers to the RABT feature in Cufflinks described in the Roberts *et al.* 2011 paper in Bioinformatics (that is not cited in the proposal). No one from the Salzberg, Irizzary or Langmead labs participated in the project (this includes the PIs). Their funding did not support the project. They are not coauthors on the paper because they did not contribute. In fact they are not even acknowledged for their help because they provided none. Even Cole Trapnell, who has participated in most of the developments to Cufflinks since its inception, played more of an advisory role (reflected in his designation as co-corresponding second to last author). The project was entirely produced by Adam Roberts with algorithmic advice from me and occasional assistance from another student of mine Harold Pimentel.

In addition to the new Cuffdiff2 module, we have made many improvements and fixed many bugs throughout the past two years in Cufflinks, Cuffdiff, Cuffcompare, and other parts of the package.

Cuffdiff2 was described in the Trapnell *et al.* 2013 paper in Nature Biotechnology. No one from the Salzberg, Irizzary of Langmead labs participated in the project. Their funding did not support the project. None of them are coauthors on the paper. Moreover, Cuffdiff2 was not a “module”. It represented an enormous effort by both the Rinn lab and my group, and especially Cole (as an example of the work that went into the paper I notes its extensive 70 page supplementary info file).

As the project has matured, we created a Google group with an email list to support both TopHat and Cufflinks jointly, because most requests for help come from users who used both tools.

This statement is misleading at best. In fact, I created the email list in January of 2011 because as corresponding author on the Cufflinks paper I was receiving a ton of email and questions, and I decided the community would be better served by more of us answering questions (I also wanted to include some of Cole’s new collaborators at Harvard/MIT, who were contributing significant new features to Cufflinks, such as CummeRbund). We have received 3516 emails asking about Cufflinks, 3494 about TopHat and 115 about CummeRbund (a newer tool for visualization of Cufflinks output developed by Loyal Goff, joint student of John Rinn and Manolis Kellis at MIT). I know these numbers because I set the password for the account when I created it.

The current project began on 7/1/2011 and runs through 4/30/2014. In just 1.8 years, it has already supported 25 peer-reviewed publications, listed in the Progress Report Publications List.

In examining this list I find reference 17:

17. A. Roberts, C. Trapnell, J. Donaghey, J.L. Rinn, and L. Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology 2011, 12:R22. PMCID: PMC3129672

This is one of the three Cufflinks publications following the first paper. Neither Steven Salzberg nor Rafael Irrizary, nor any of their students or group members were involved at all in this project. Moreover, they were not funding any of the personnel on the project at the time the work was done. This is blatant plagiarism of work they had nothing to do with. Moreover, I noticed that reference 16 on the list from the Rinn lab also appears to be gratuitously added, the only connection to Salzberg or Irizzary that I can imagine being the fact that the *fifth* author was a *former* student of Salzberg.

16. T.R. Mercer, D.J. Gerhardt, M.E. Dinger, J. Crawford, C. Trapnell, J.A. Jeddeloh, J.S. Mattick, and J.L. Rinn. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nature Biotechnology 2011 30(1):99-104.

Allele-specific variation. RNA-seq data contains all the information required to study allele-specific expression (associations between expression level and allele). In fact, this is a very commonly requested feature for TopHat and Cufflinks. We propose to build such a feature into TopHat and Cufflinks

It is true that this has been a frequently requested feature. For this reason, I submitted it as a goal in my R01 grant (R01 HG006129, Aim 1e). My group released a solution to this problem by optimizing our new program eXpress (that addresses goals in Aim 2 of my grant) for allele specific expression. Secondly, Cole is about to release an allele specific expression solution for Cufflinks (work that has resulted from a collaboration he has undertaken with Jun Liu at Harvard). It is true that a PI would not know about our aims or work and could, in principle, propose such a feature independently in a grant application. But Steven Salzberg did know of this work. I quote from the beginning of an email of his from the time when we were coordinating submission of our two proposal submission (his HG006102 and my HG006129) in 2010:

```
Steven Salzberg <salzberg@umiacs.umd.edu> Fri, Apr 2, 2010 at 2:10 PM
To: lpachter@math.berkeley.edu
Cc: Steven Salzberg <salzberg@umd.edu>
hi Lior,
I am almost done with this proposal.  I removed part of Aim 3, which is
the Cufflinks part, dealing
with allele specific expression - so that you can
include that in your grant.
```

We have augmented this informal support with continuing updates to our manual pages and with a published protocol [4], all of which give users step-by- step instructions on how to run the most common types of RNA-seq analyses.

and also

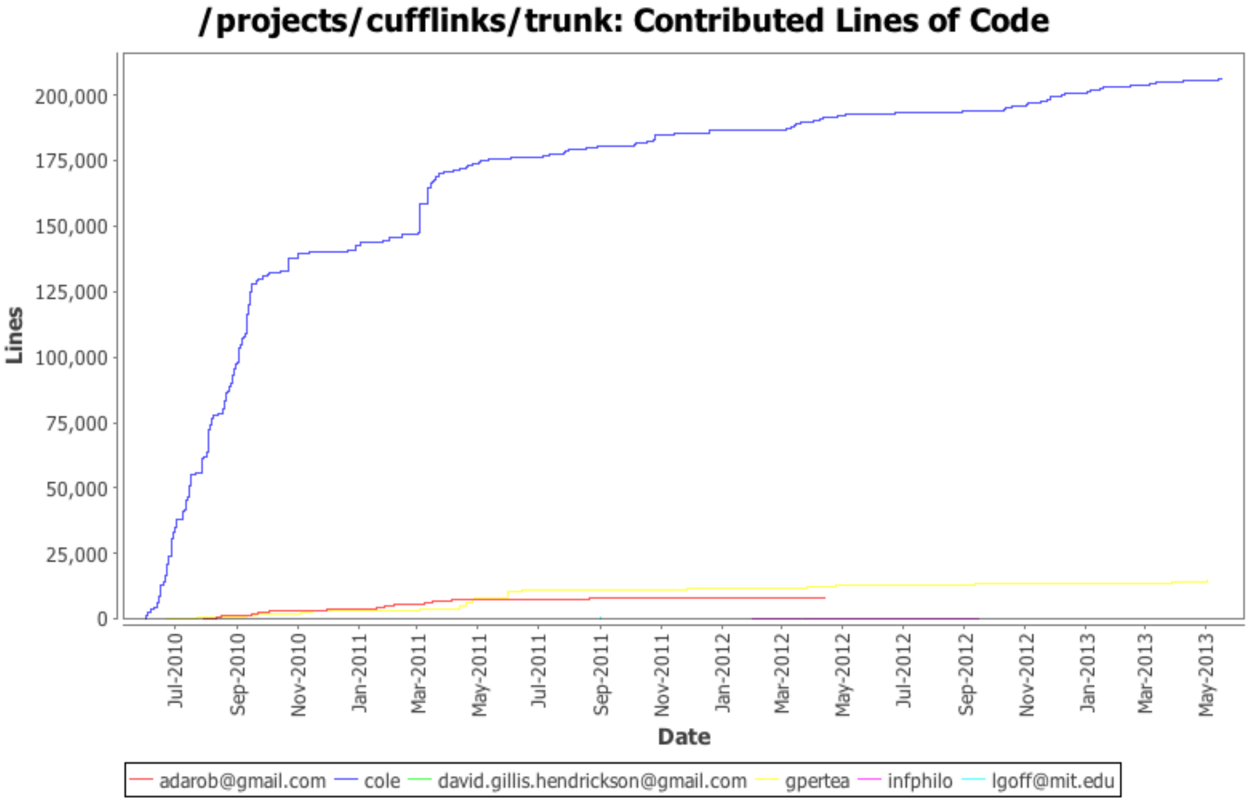
...we described how to use Cufflinks (and Bowtie/TopHat) for this scenario in our Nature Protocols paper in 2012 [4].

The paper [4] refers to:

- C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn and L. Pachter, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Nature Protocols, 7 (2012), 562–578.

This paper describes the overall Tuxedo pipeline for RNA-Seq and was published last year. Steven Salzberg is the third to last author, and his two students who have supported TopHat/Cufflinks are 4th and 5th authors. There is justification for this author placement. Although Salzberg's group has been helpful in answering questions on the email list, and in fixing occasional bugs, and even though Daewhan Kim has played a major role in the development of TopHat2 which forms a part of the pipeline, it is Cole who developed the bulk of Cufflinks, Adam Roberts (my student) who took over and was primary developer after Cole left my group for his postdoc, and Loyal Goff (student of John Rinn and Manolis Kellis) who undertook a major project in developing CummeRbund which is the primary means now for users to interact with Cufflinks. Thus it is untrue to claim that "we have augmented informal support" or to present [4] as "our Nature Protocols paper". In terms of support for Cufflinks, it is not even true that Salzberg's group is the primary group helping users via the email list or in other ways. Not only does my group answer Cufflinks email, last year I hosted a workshop/conference at UC Berkeley, called *Seq I to provide tutorials to users on RNA-Seq analysis, specifically with Cufflinks (as promised in my grant). It was attended by almost 200 people (with many viewing later by video). Loyal Goff and Cole Trapnell presented along with my students Harold Pimentel and Adam Roberts.

To make concrete the relative contributions to Cufflinks software, I note that Cole Trapnell has written 90% of the code for the Cufflinks project (as determined from an analysis of the code repository, see Figure below). His contribution is also evident in having made 74% of the changes since the software was written. Also note that the majority of the code was written *after* Cole graduated in the summer of 2010. My student Adam Roberts has made 16% of the changes, in comparison to Geo Pertea (Salzberg group) who has made only 9% of the changes. However contribution to a project cannot be measured on the basis of lines of code, or changes committed, alone (any more than the impact of a paper can be purely measured in terms of citations). The fact is that Geo Pertea's contributions were to Cuffcompare and to the GTF (annotation) parser. These are auxillary tools to the main Cufflinks software that, though important for the functioning of the program, involve none of the complex statistical, mathematical and algorithmic ideas developed by Cole Trapnell, Adam Roberts, and myself, that have made Cufflinks a success.



Contributions to the Cufflinks project: lines of code written.

Afterword

I was relieved to hear from you during our phone conversation that NIH will take my allegations very seriously. I do not make them lightly and I realize the gravity of my claims. I have been on good terms with Steven Salzberg for many years now, and although we have not collaborated we have been regularly helping each other with the Tuxedo suite. Geo Pertea, in Salzberg's lab, has been answering questions asked on our email list, and has made bug fixes when needed and responded promptly to our occasional requests for tweaks to Cuffcompare. Conversely, my student Harold Pimentel has assisted on TopHat, to the extent that he implemented a feature for TopHat2 and is a co-author on the paper (although I am not a coauthor, I am specifically acknowledged, together with my student Adam Roberts, for our help on TopHat2). Similarly, I respect and have been on good terms with Rafael Irrizary, and also Ben Langmead. It therefore pains me to make accusations of conduct regarding the submission of the HG006102 proposal and of plagiarism of my work. However I must do so because it is evident to me that in the submission of the renewal for HG006102 the PIs have deliberately taken credit for 3 years of work on Cufflinks that is not theirs. I have tried my best to give the PIs the benefit of doubt, but I am convinced that any reader of the proposal would consider it a forgone conclusion that Steven Salzberg (possibly with Rafael Irizzary) developed the Cufflinks project and that Steven Salzberg together with Ben Langmead are the best team going forward. In the Summary section of the proposal the PIs write:

The papers describing these systems [Bowtie, Tophat and Cufflinks] already have thousands of citations.. We will continue to develop [the software tools] ... with our small team.. One reason we are able to support so many users with such a small group is that the genomics community has developed popular web resources and user forums through which many other scientists and engineers answer questions and provide suggestions for new users. We will continue to support these external sites and encourage others to use and if possible modify and enhance our software.

In fact, the Cufflinks project has been successful because there are many of us who have worked very hard on it. Not only students in my group but also in the Rinn Lab. Together at least 10 people from our groups have made major contributions, in ideas, software development and experimentation. Our combined work in writing three high impact papers on Cufflinks following the initial publication was not a triviality. And it goes without saying that Cole Trapnell is primarily responsible for the success of the project thanks to his talents as a scientist, his extraordinary software engineering skills, his biological intuition and his immense effort. In summary, the meaning of "free" when I make my tools and software freely available under unrestrictive open source licenses is that the community can adopt the tools without barrier. "Free" is not a license for plagiarism by colleagues.

Sincerely,

Lior Pachter