1. Research Programs Engaging the Research Specialist

Unit Director Aviv Regev engages in many cancer research programs to which Research Specialist Brian Haas contributes. Most notably, Regev is the principal investigator of the **NCI-funded U24 grant (1U24CA180922-01) titled "Trinity: Transcriptome assembly for genetic and functional analysis of cancer."** Haas co-developed Trinity for *de novo* transcriptome assembly using RNA-Seq data (Grabherr, Haas et al. 2011); has expanded it to include a toolkit providing support for downstream investigation of transcript structure, function, and expression (Haas, Papanicolaou et al. 2013), and is the lead of the Trinity team. Trinity is regularly used by thousands worldwide, and has been cited in >2500 papers since 2011, >700 in cancer. The project led by Haas under the U24 funding is committed to leveraging RNA-seq assembly and data in the Trinity package as a cutting edge and robust tool for the unique challenges and opportunities of the cancer transcriptome.

In particular, the U24 grant supports **continued development of the Trinity software as a leading Cancer Transcriptome Analysis Toolkit (CTAT)**. Its aims include **(1) enhancing the Trinity analysis modules for understanding cancer transcriptomes and genomes** by incorporating new methods and software for the detection of cancer mutations, fusion transcripts, RNA-editing, lincRNA identification, reconstruction of transcripts not represented by the reference human genome such as derived from viruses or microbes, and support for characterizing the heterogeneity of tumors via single cell transcriptome analysis; **(2) maintaining and updating Trinity to leverage emerging technologies and tools,** such as long-read sequence technology derived from PacBio or Moleculo and modern Bioconductor packages for studying differential expression; **(3) providing optimized and efficient Trinity applications for cancer researchers in diverse environments**, making the tools available for users to either install and execute in their own environments or execute via the Galaxy web portal at collaborating institution Indiana University; and (**4) growing, training, and supporting a Trinity community of cancer researchers** through hands-on workshops, training videos, and extensive on-line documentation. These aims -- and the additional cancer research activities to which Unit Director Aviv Regev and Research Specialist Brian Haas contribute -- are further detailed below. **We note that Haas is both the key person in all of these activities, and without his leadership and his hand-on work, Trinity would have neither existed, nor have been adopted by thousands, nor have been extended now to specific cancer applications.**

1.1. Enhancing Trinity analysis modules for understanding cancer transcriptomes and genomes.

RNA-Seq coupled with genome-free assembly provides a direct gateway to the expressed portion of the cancer genome. To this end, we have been enhancing Trinity's software modules to leverage transcriptome assembly to detect SNPs in expressed exons and introns, identify cancer fusion transcripts, detect evidence of transcript editing and alternative splicing, identify viral or microbial transcripts, and to determine tumor heterogeneity from single cell transcriptomes. In each effort, we are demonstrating the utility of each module through **Driving Cancer Projects**.

For ***variant detection***, we have been pursuing multiple strategies based on RNA-Seq read mappings to reference genomes and transcript assembly-based methods, and through collaboration with the Broad Genome Analysis Toolkit (GATK) (McKenna, Hanna et al. 2010, Van der Auwera, Carneiro et al. 2013), we have recently made available GATK-based variant detection from RNA-Seq as part of our Trinity CTAT. We are actively pursuing advanced methods for variant detection based on *de novo* transcriptome assembly.

To enable ***fusion transcript discovery***, we have recently developed multiple tools that leverage Trinity and that are now incorporated into Trinity CTAT, including STAR-Fusion (http://star-fusion.github.io), DISCASM (http://discasm.github.io), and FusionInspector (http://fusioninspector.github.io). These tools highlight the evidence for fusion transcripts as found within the RNA-Seq reads, and showcase Trinity *de novo* reconstructed fusion transcripts, coupled with identification and annotation of fusions previously associated with tumorigenesis. Haas is now working with the Broad CLIA-certified lab (CRSP), to use these as a backbone for fusion transcript analysis in patient samples in clinical care (**Section 1.7**).

For ***lincRNA classification,*** efforts are underway to integrate into Trinity CTAT our newly developed SLNCKY software (http://slncky.github.io/) (Chen, Shishkin et al. 2016).

For ***tumor heterogeneity***, our group has made great strides in characterizing the heterogeneity of patient tumor samples via single cell RNA-Seq (see below). We are currently incorporating new algorithms and software into Trinity CTAT to support future single cell tumor transcriptome studies.

1.2. Maintaining and updating the Trinity software to leverage emerging technologies and tools.

Sequencing technologies continue to advance at a staggering pace, and we have been pursing efforts to ensure that Trinity will evolve to maintain compatibility with and exploit the advantages available from the latest advances in transcriptome sequencing. For example, we have recently reengineered the Trinity assembler to leverage long transcript reads as generated by PacBio instruments, and we are actively pursuing *in silico* experiments to assess the advantages of long PacBio reads, long Illumina paired-ends, and combinations of technologies in Trinity *de novo* assembly. We have also participated in the Oxford Nanopore MinION Early Access Program and have been sequencing transcriptomes via its new MinION device. While there are challenges in currently using these nanopore data for Trinity assembly due to sequencing error rates and error modes, we are committed to leveraging and supporting the technology where we can demonstrate its added value in shedding light on cancer transcriptomes.

1.3. Providing optimized and efficient Trinity applications for cancer researchers in diverse environments

Trinity CTAT includes several software tools and data resources that require considerable computational resources, both in terms of data storage and computing hardware requirements. Trinity is Open-Source and all Trinity source code and related data resources are made freely and openly available. However, researchers with varying informatics skills and access to high-performance computing infrastructures will also have varying abilities to leverage all that Trinity CTAT provides. We therefore offer differing modes of execution: (**1**) a Trinity CTAT Galaxy Web Portal (https://galaxy.ncgas-trinity.indiana.edu) to assist cancer researchers who simply want to upload their RNA-Seq data and use an intuitive web browser to execute computes and analyze their data; and (**2**) access to code and documentation for independent execution by expert bioinformaticians. We are also actively engaged with additional groups to provide software and computation specifically to cancer researchers, including the **NCI Cloud project** and **GenomeSpace** (see below), which will further broaden the reach of Trinity CTAT within the cancer research community.

1.4. Growing, training, and supporting a Trinity user community of cancer researchers

Our Trinity online documentation (http://trinityrnaseq.github.io) provides a resource for those pursuing RNA-Seq studies, and highlights our Trinity CTAT activities. We make Trinity workshop materials available in the form of user guides, videos, and virtual machines that contain all relevant software and data to support learning activities. We also provide RNA-Seq workshops periodically in conferences throughout each year, involving a combination of lecture and hands-on exploration of RNA-Seq data using our toolkit. We directly engage with Trinity users via our Google forum, which has several hundred posts each month; the user community is increasingly playing a key role in assisting other users with technical support.

1.5. Characterizing tumor cell heterogeneity

The Regev Group has also been a pioneer and leader of single cell RNA-Seq and its application to fresh human tumors. Recently published work by the Regev group used single cell transcriptomes to study primary glioblastoma and demonstrated the ability to detect mutations, copy number variation including the loss and gain of individual chromosomes or chromosome segments, and oncogenic transcript isoforms; it was also able to classify cells according to expression signatures of tumor types (Patel, Tirosh et al. 2014). Over the past 2 years, the lab has worked with the Dana Farber Center for Cancer Precision Medicine (CCPM) to develop a robust experimental and computational pipeline for single cell RNA-Seq of patient tumors, including resections, core biopsies, FNAs, ascites and pleural effusions across many different tumor types. Algorithms and tools developed as part of these efforts are being adopted and integrated by Haas' team into the Trinity CTAT for broader accessibility to the cancer research community.

1.6. Large-scale computational analysis of cancer data

Dr. Regev's collaborative efforts extend to other large cancer research initiatives at the Broad Institute for which Haas contributes, including The Cancer Genome Atlas (**TCGA**) (Cancer Genome Atlas Research, Weinstein et al. 2013), the **NCI Cloud project**, and **GenomeSpace** (http://www.genomespace.org/), a popular framework for integrative bioinformatics data analysis (Qu, Garamszegi et al. 2016).

TCGA and the NCI Cloud Pilot. TCGA applies genome analysis technologies including large-scale biological sequencing to characterize the molecular basis of cancer. The **NCI Cloud Pilot** project aims to provide researchers with access to the massive volumes of data generated by TCGA and to compute resources and tools to effectively interrogate these data. Broad Institute investigator Gad Getz is a co-PI of the Genome Data Analysis Center of the NCI/NHGRI TCGA project and co-PI of the Broad Institute's NCI Cloud Pilot 'Firecloud' (**Gad Getz, recommendation letter**). Broad Chief Data Offier and physician-scientist Anthony Philippakis is leading the development of our computational infrastructure to support the NCI Cloud Pilot (**Anthony Philipakkis, recommendation letter**). Unit Director Regev routinely collaborates with Getz and Philippakis in sharing computational algorithms, software tools, and laboratory technological developments, along with co-advising postdoctoral scientists. Efforts are underway to integrate the Trinity CTAT into Firecloud so that computational results may be made available as companion data sets in TCGA, and to enable users of Firecloud to have access to Trinity in analyzing additional cancer data sets in using this cloud-computing platform. Additional opportunities are available for incorporating Trinity CTAT into the NCI-funded cloud-computing platform at Seven Bridges (**Lu Zhang, recommendation letter**).

GenomeSpace. Another route for the cancer research community into Trinity CTAT is GenomeSpace, a cloud-based framework that supports interoperability among a diverse range of bioinformatics tools, supporting analysis through an easy-to-use web interface. Director Regev is co-PI for the GenomeSpace project, along with co-PI Jill Mesirov (previously at the Broad Institute, currently the Associate Vice Chancellor for Computational Health Sciences at the UC San Diego School of Medicine). Integration of Trinity CTAT into the GenomeSpace ecosystem of tools is currently underway (**Jill Mesirov, recommendation letter**).

1.7. Translating methods and knowledge to clinical applications

Ultimately, we aim to translate scientific knowledge and bioinformatics methods into clinical applications in order to improve patient care and accelerate diagnostics and treatments for cancer patients. To advance this effort, the Broad Institute recently launched a Clinical Research Sequencing Platform (CRSP) to provide highly accurate genome sequencing for clinicians treating patients either as routine care or in the context of clinical trials. Niall Lennon directs the clinical development of CRSP, and strongly supports our proposal to introduce Trinity CTAT modules for **clinical sequence analysis** (**Niall Lennon, recommendation letter**). Initial applications of the Trinity toolkit to Chronic Lymphocytic Leukemia (CLL) patient samples, in collaboration with Catherine Wu of Dana Farber, have proven useful in identifying recurrent fusion transcripts that may be highly relevant to CLL etiology (**Catherine Wu, recommendation letter**). Through continued collaborative efforts with the CRSP, we expect to have a measurable positive impact on cancer research and clinical care (**Niall Lennon, recommendation letter**).

2. Role of the Research Specialist in Research Programs

Research Specialist Brian Haas plays many key roles in the above programs, involving computational scientific research, software and algorithm development, project leadership and collaboration, community support and training, and manuscript development for publication. Indeed, Haas is **the key person for Trinity**, leads its team, does a substantial amount of the method development, software development and outreach, and his contribution is thus **singularly critical**. Each specific role is outlined below.

2.1. Computational scientific research

Cancer is a disease of the genome, and studying and understanding it as such requires a high level of proficiency in genome sequence exploration and analysis. Haas's entire professional career (since 1999) has involved many aspects of genome research, spanning annotation and analysis of model organisms, disease-causing pathogens, insect vectors of disease, the human microbiome, and most recently cancer genomes and cancer transcriptomes.

Much of Haas' earlier research efforts have involved studies of transcriptomes and genomes. Transcriptome-focused research efforts include identification of micro-exons (Volfovsky, Haas et al. 2003), the use of full-length cDNAs and ESTs to study alternative splicing (Haas, Volfovsky et al. 2002, Haas, Delcher et al. 2003, Campbell, Haas et al. 2006), identification of sequence signals for polyadenylation site recognition (Loke, Stahlberg et al. 2005, Shen, Ji et al. 2008), exploring the depth of RNA-Seq coverage required for comprehensive transcriptome investigations (Haas, Chin et al. 2012), ribosome profiling to identify novel coding regions of expressed transcripts (Fields, Rodriguez et al. 2015), and studies of variation among single cell transcriptomes (Kowalczyk, Tirosh et al. 2015).

The genome research activities that Haas has played pivotal roles in are extensive and include studies of repetitive DNA structure (Volfovsky, Haas et al. 2001), segmental genome duplications (Haas, Delcher et al. 2004), genome synteny (Nene, Wortman et al. 2007), paralogous family clustering (Haas, Wortman et al. 2005), loss of heterozygosity (Jiang, de Bruijn et al. 2013), and gene family evolution (Haas, Kamoun et al. 2009).

In Haas' current work, research involves exploring the accuracy of fusion transcript detection in cancer, coupled with the development of novel methods to advance this area. This includes developing synthetic and genuine RNA-Seq data sets to benchmark the accuracy of existing and newly developed methods, and applying methods to detect novel cancer fusion transcripts in samples from cancer cell lines and from clinical patient samples.

All research activities have required Haas to be highly proficient in leveraging available bioinformatics tools, using popular data resources (eg. GenBank, Ensembl, etc.), and in being able to develop new software tools as needed (see Section 2.2 below). Haas' combined training in both molecular biology and computer science, as well as his extensive experience in bioinformatics and genome sequence analysis, is critical in gaining further insights into cancer genomes as genomic data continue to accumulate. Going forward, Haas is expected to make essential contributions to research activities exploring cancer transcriptome data made available by TCGA (**Gad Getz, recommendation letter**), from clinical patient samples (**recommendation letters from Niall Lennon and Catherine Wu),** and to studies of tumor heterogeneity via single cell transcriptome analysis (**Aviv Regev, recommendation letter**).

2.2. Computational Methods and Software Development

Biological research, computational methods development, and software development are often heavily intertwined -- but frequently, existing computational methods are not perfectly suited for genome research; they may for instance lack appropriate speed, accuracy, or usability. Haas has **repeatedly advanced the field** by developing new methods to tackle such difficult challenges in genome analysis. Examples apply to every major category of genome analysis projects. Genome annotation and analysis efforts were greatly facilitated by his PASA (Haas, Delcher et al. 2003) and EVidenceModeler (Haas, Salzberg et al. 2008) software. PASA aligns transcript sequences to a reference genome and assembles transcripts into maximal alignment assemblies, followed by incorporating these alignment assemblies into gene predictions, refining exon boundaries, adding untranslated region (UTR) annotations, and modeling alternatively spliced isoforms. EVidenceModeler assists with genome annotation by incorporating evidence derived from protein alignments, transcript alignments, and ab initio gene predictions into weighted consensus gene structure annotations. PASA and EVidenceModeler are used together as part of a robust genome annotation framework (Haas, Zeng et al. 2011), used by the Broad Institute production genome annotation group and other research efforts around the globe, by institutes (eg. Joint Genome Institute and the J. Craig Venter Institute) as well as by individual researchers.

Haas also developed a suite of tools to help study the human microbiome, focused on 16S rRNA sequence analysis to study microbiome complexity and to detect artifacts resulting from chimeric reads (Haas, Gevers et al. 2011). The microbiome utilities toolkit included a new algorithm for 16S rRNA chimera detection, fixed-width 16S rRNA sequence alignment, and operational taxonomic unit clustering. Algorithms were later adapted into other popular toolkits including Mothur (Schloss, Westcott et al. 2009) and QIIME (Caporaso, Kuczynski et al. 2010) and used extensively as part of the human microbiome analysis effort (Human Microbiome Project 2012).

Most recently, Haas co-developed the popular **Trinity *de novo* RNA-Seq assembly software** (Grabherr, Haas et al. 2011, Haas, Papanicolaou et al. 2013), and is leading its ongoing development and application to cancer transcriptomes as part of our more comprehensive Trinity CTAT. These activities have leveraged Haas' knowledge and expertise in software development and bioinformatics. Trinity CTAT has code written in several computing languages, including Perl, Python, C++, Java, and R. The software makes use of high performance computing grid architectures, including LSF and SGE. Relational databases and web-based user interfaces leveraged by the system require knowledge of SQL, HTML and Javascript. Haas' capabilities in each of these areas have enabled productive development of software that is widely used and appreciated by a broad user community.

During the development of Trinity CTAT, we concluded that important challenges remained for the fast and accurate detection of cancer fusion transcripts. In collaboration with Alex Dobin, Haas developed several new tools that are now incorporated into Trinity CTAT, including DISCASM, STAR-Fusion, and FusionInspector (Alex Dobin, recommendation letter; manuscript in prep.), providing unparalleled speed and accuracy for fusion detection. The tools are now readily available to all cancer researchers via our Galaxy portal's point-and-click web interface, and Haas is working with CRSP (**section 1.7**) to bring these to CLIA-certified clinical sequencing.

2.3. Open source software, community access, and community-assisted development

To best serve the cancer scientific community, software for genomic analysis should be readily available, well documented, and straightforward to use. Each software tool developed by (or in collaboration with) Haas is open source and has been made widely available and accessible to users via popular outlets including sourceforge.net and github.com (**recommendation letters from Steven Salzberg and John Quackenbush**). All tools have been well documented and supported by Haas throughout their existence. One of the earliest such tools, PASA (originally released in 2003), continues to be popular among genome researchers and genome centers, and is accordingly well maintained.
The Trinity software was published and released as open source in 2011, and has since been downloaded >50,000 times. Because it is open source, many community developers have studied the code and contributed back changes to improve execution efficiency. Haas has worked closely with members of the community to integrate their changes into the Trinity codebase. This has increased the software's popularity and uptake by other cancer research groups across the globe (**Chris Mason, Francis Ouelette, and Remy Bruggmann recommendation letters**), and has garnered interest in both academia and industry (e.g., **recommendation letter from Ben Zeskind, Rebecca Kusko, and Maxim Artyomov of Immuneering, Inc., and recommendation letter from Lu Zhang of Seven Bridges, Inc.**).

Haas remains committed to ensuring that researchers have ready access to the software and algorithms we develop to further support cancer research. Haas' deep level of commitment and mission-driven work is critical for the project's success and deeply impactful across the community.

2.4. Project collaboration and leadership

Effective project leadership and productive collaborations are a key to success in science, particularly in the cancer research programs described above. Haas has a long track record of successful scientific collaborations as demonstrated by his extensive list of co-authored publications (see biosketch and **Section 2.6**). His leadership roles in scientific projects include his earlier role at the Broad Institute as Manager of Collaborations, Outreach, Bioinformatics, Research and Analysis (COBRA), where he managed a group of bioinformatics analysts and engineers supporting numerous collaborative genome analysis projects.

In his current role as Senior Computational Biologist, in addition to the research and development he performs himself, he manages a small group of software engineers who support development and computational infrastructure for the Klarman Cell Observatory cancer research activities. Haas also provides project leadership for Trinity CTAT in collaboration with partners at Indiana University (**Tom Doak, recommendation letter**).

Future success of Haas' role in cancer research projects will require effective collaboration with the TCGA research efforts led by Gad Getz, the Klarman Cell Observatory led by Aviv Regev, and Haas' continued leadership for the Trinity CTAT project.

2.5. Community support and training

Software to facilitate cancer research will be widely adopted only if its users are well-supported and given proper training in how to best make use of the tools and related resources. Haas has explored many avenues to support the Trinity CTAT user community, including extensive documentation on the Trinity website (https://github.com/trinityrnaseq/trinityrnaseq/wiki), YouTube videos to describe the toolkit and related efforts (https://www.youtube.com/watch?feature=player_embedded&v=9ky5NwV45qY, and http://www.broadinstitute.org/partnerships/education/broade/trinity-screencast) , an active Google forum for users (https://groups.google.com/forum/#!forum/trinityrnaseq-users), a Twitter feed with Trinity CTAT announcements and Galaxy portal service updates (https://twitter.com/Trinity_CTAT), and hands-on training workshops on using Trinity for RNA-Seq studies (https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/wiki).

Haas is expected to continue user support and training activities in the cancer research programs. This includes attracting researchers to the NCI Cloud, engaging with users via forums, and providing training and support for Firecloud RNA-Seq related cancer research efforts, particularly those involving Trinity CTAT. In addition, Haas would play an active role in the community by advocating the use of the general technologies developed to support scientific workflow computation underpinning the NCI Cloud (**Anthony Philippakis, recommendation letter**), enhancing the impact of these technologies and NCI Cloud efforts.

2.6. Manuscript development

Haas' extensive contributions to genome annotation, analysis, and bioinformatics methods and software development have been published in top scientific journals including **first author papers in *Nature*, *Nature Biotechnology*, and *Genome Research*** (see Biosketch). In general, his publications are highly cited; for example, the *Nature Biotechnology* publication of Trinity (Grabherr, Haas et al. 2011) has been cited >2,500 times, and the *Nature Protocols* paper describing the Trinity analysis protocol (Haas, Papanicolaou et al. 2013) has been cited ~500 times. On multiple occasions, including last year (2015), **Haas was named by Thomson Reuters as a Highly Cited Researcher**. In another honor*, Nature Biotechnology* plans to highlight the 2011 Trinity publication in its upcoming celebratory 20th anniversary retrospective issue. Haas is expected to continue to contribute to manuscripts that describe advances in cancer research.