UNIVERSITY OF CALIFORNIA SAN DIEGO

# Building bioinformatic tools for massive repurposing of multi-omic data in the Sequence Read Archive

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Brian Yik Tak Tsui

Committee in charge:

Professor Hannah Carter, Chair
Professor Jill Mesirov, Co-Chair
Professor Ruben Abagyan
Professor Terry Gaasterland
Professor Nathan Lewis

2019

The Dissertation of Brian Yik Tak Tsui is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

---

---

---

---

---

University of California San Diego

2019

# DEDICATION

I want to thank all my friends and family members who have made my Ph.D. journey fun and enjoyable.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

First and foremost, I would like to thank my primary advisor, Dr. Hannah Carter, for giving me the opportunity to work with her over the past four and a half years. Though I had little research experience when I finished my undergraduate study, she chose to take me under her wing and guided me through the ups-and-downs of research. Hannah is a role model as a faculty member and also as a leader. Hannah exhibits the highest standard as both a scientific researcher and also as a mentor. Her seriousness and dedication to science and work have become a significant milestone in my professional character building. I would also like to thank her unyielding faithfulness to our mentor relationships even at times when I disappointed her repeatedly.

I would also like to thank my dissertation committee members for their guidance and support for the last two years. Also, I would like to thank Dr. Trey Ideker who generously shared his computational resources and his expertise to make my thesis work possible. Also, I would like to thank my lab members. Specifically, I would like to acknowledge Rachel Marty, Michelle Dow, Kivil Ozturk, Macarena Fernandez Chacon, Andrea Castro and Jessica Zhou for accompanying me through the last couple of years. I would also like to thank Dylan Skola and Shamim Mollah for their supports as friends in my scientific journey. I would like to finally thank Dr. Dean Tullsen for mentoring me during undergraduate research, who took me under his wing along with the associated lab members Dr. Hueng Wei Tseng and Dr. Ashish Venakat who have helped me. The level of mentor care that Dr. Dean Tullsen offers to students is something that I can never forget.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Extracting allelic read counts from 250,000 human sequencing runs" in *Pacific Symposium on Biocomputing*, 2019 by Brian Tsui, Michelle Dow, Dylan Skola, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is a reformatted presentation of the material currently being prepared for submission for publication as "SkyMap JupyterHub: A cloud platform to query and analyze >900,000 re-processed public sequencing experiments from >20,000 studies" by Brian Tsui, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is being prepared for submission, is a reformatted reprint of the material as it appears as "Creating a scalable deep learning based Named Entity Recognition Model for biomedical textual data by repurposing BioSample free-text annotations" in *bioRXiv*, 2018 by Brian Tsui, Shamim Mollah, Dylan Skola, Michelle Dow, Chun-Nan Hsu, Hannah Carter. The dissertation author was the primary investigator and author of this paper.

# VITA

2014 University of California San Diego
*Bachelor of Science, Computer Science*

2019 University of California San Diego
*Doctor of Philosophy, Bioinformatics and Systems Biology*

PUBLICATIONS

**Brian Y Tsui,** Michelle Dow, Dylan Skola, Hannah Carter. 2019. Extracting allelic read counts from 250,000 human sequencing runs in Sequence Read Archive. 196-207. 10.1142/9789813279827_0018. *Proceedings of the Pacific Symposium*.

**Brian Y Tsui,** Shamim Mollah, Dylan Skola, Michelle Dow, Chun-Nan Hsu, and Hannah Carter. 2018. "Creating a Scalable Deep Learning Based Named Entity Recognition Model for Biomedical Textual Data by Repurposing BioSample Free-Text Annotations." *bioRxiv*. https://doi.org/10.1101/414136.

**Brian Y Tsui,** Michelle Dow, Dylan Skola, Hannah Carter. 2019. "SkyMap JupyterHub: A cloud platform to query and analyze >900,000 re-processed public sequencing experiments from >20,000 studies" *Manuscript in preparation.*

Huiying Liang*, **Brian Y Tsui***, Dr. Carolina Carvalho Soares Valentim*, Hao Ni*, Sally Baxter*, Guangjian Liu , Jiancong Chen , Mr. Daniel Kermany , Xin Sun , Liya He , Jie Zhu , Dr. Wenjia Cai , Adriana Tremoulet , Lianghong Zheng , Rui Hou , Gen Li , Ping Liang , Xuan Zang , Zhiqi Zhang , Liyan Pan , Huimin Cai , Rujuan Ling , Shuhua Li , Yongwang Cui , Shusheng Tang , Hong Ye , Xiaoyan Huang , Waner He , Wenqing Liang , Qing Zhang , Jianmin Jiang , Wei Yu , Jianqun Gao , Wanxing Ou , Yingmin Den , Qiaozhen Hou , Bei Wang , Cuichan Yao , Yan Liang , Shu Zhang , Mr. Yaou Duan , Runze Zhang , Sierra Hewett , Charlotte Zhang , Oulan Li , Michael Goldbaum , Hannah Carter , Long Zhu , Huimin Xia , Sarah Gibson , Gabriel Karin , Xiaokang Wu , Nathan Nugyen , Cindy Wen , Jie Xu , Dr. Chun-Nan Hsu , Dr. Wenqin Xu , Bochu Wang , Pin Tian , Hua Shao , Edward Zhang , Winston Wang , Jing Li , Bianca Pizzato , Caroline Bao , Daoman Xiang , Wanting He , Suiqin He , Yugui Zhou , Weldon Haw. "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence" *Nature Medicine*, 2019.

Dow, Michelle, Rachel M. Pyke, **Brian Y Tsui**, Ludmil B. Alexandrov, Hayato Nakagawa, Koji Taniguchi, Ekihiro Seki, Olivier Harismendy, Shabnam Shalapour, Michael Karin, Hannah Carter, and Joan Font-Burgada. 2018. "Integrative Genomic Analysis of Mouse and Human Hepatocellular Carcinoma." *Proceedings of the National Academy of Sciences of the United States of America 115* (42): E9879–88.

Daniel Ortiz, **Brian Y Tsui**, Tyler Goshia, Colleen Chute, Anthony Han, Hannah Carter, and Stephanie I. Fraley. 2017. "3D Collagen Architecture Induces a Conserved Migratory and Transcriptional Response Linked to Vasculogenic Mimicry." *Nature Communications* 8 (1): 1651.

Wang, Tina, **Brian Y Tsui**, Jason F. Kreisberg, Neil A. Robertson, Andrew M. Gross, Michael Ku Yu, Hannah Carter, Holly M. Brown-Borg, Peter D. Adams, and Trey Ideker. 2017. "Epigenetic

Aging Signatures in Mice Livers Are Slowed by Dwarfism, Calorie Restriction and Rapamycin Treatment." *Genome Biology* 18 (1): 57.

Gonzalo Almanza, Jeffrey J Rodvold, **Brian Y Tsui**, Kristen Jepsen, Hannah Carter, Maurizio Zanetti. 2018. "Extracellular vesicles produced in B cells deliver tumor suppressor miR-335 to breast cancer cells disrupting oncogenic programming in vitro and in vivo." Scientific reports 8 (1): 17581.

*These authors contributed equally to this work.

# ABSTRACT OF THE DISSERTATION

Building bioinformatic tools for massive repurposing of multi-omic data in the Sequence Read Archive

by

Brian Yik Tak Tsui

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2019

Professor Hannah Carter, Chair
Professor Jill Mesirov, Co-Chair

There are currently millions of sequencing experiments generated from various research groups. Each study often offers a pre-processed -omic matrix as a supplementary table in the manuscript or is deposited to public data repositories. The pre-processed -omic data is essential to the research community to derive -omic based statistical and machine learning predictive models

which can be useful for understanding biological features. While the analysis-ready pre-processed -omic data minimizes the prerequisite of high-performance computing and data processing knowledge for researchers, the lack of consistency in these submitted preprocessed data and their associated metadata pose a challenge towards secondary data analysis.

Towards the goal of tackling this challenge, the first chapter of the thesis evaluates the possibility of reprocessing and extracting the allelic read counts from over 250,000 sequencing experiments and 10,000 public studies in the SRA, which is useful towards identifying novel variant associations and evaluating the allele-specific expression. The second chapter assesses the possibility of constructing an online platform in which the research community can smoothly to go from data querying to analysis without any programming background. The third chapter tackles the metadata aspect of the SRA by evaluating the possibility of recognizing the biomedical entities in the metadata without expert curation. Through repurposing the vast amount of submitter-based biospecimen annotations, we can train a deep-learning-based model to evaluate the relationships between various annotations. In summary, this dissertation builds the foundation towards automating the process of -omic data association with metadata.

# INTRODUCTION

The central dogma of biology was first discovered during the 1970s and is stated as the following: cells contain deoxyribonucleic acid (DNA), which is often then transcribed into ribonucleic acid (RNA). The RNA is then translated into amino acids and folded into proteins (Crick, 1970). However, it is only in recent years where the advance of sequencing technology has dramatically improved our ability to capture the DNA genomic sequences and RNA transcript sequences of the cells in wholeness, and the word "wholeness" has Greek root "-omic" (Yadav, 2007), hence are called the gen-omic data and transcript-omic data respectively. Moreover, this vast amount of -omic data collected by many public studies is increasing both in volume, and in diversity of library strategies and experimental conditions.

However, this vast amount public -omic data is in large part under-explored because it is stored in different formats or processed by different bioinformatics pipelines. As a result, the data have poor Findability, Accessibility, Interoperability, and Reusability (FAIR) (Wilkinson et al., 2016). One fundamental problem is that the existing bioinformatics pipelines usually offer statistics as the output. For example, variant callings and identification of differentially expressed genes are often done with statistical-based tools like Mutect (Cibulskis et al. 2013) and DE-seq (Love et al. 2014). However, many of these statistical models require assumptions that might or might not fit the data and it is difficult to evaluate the statistical biases in over 10,000 sequencing studies. This situation requires a better solution.

This dissertation leverages the idea that both science and statistics rely on counting. Siddhartha famously argued that "Science starts with counting" in his Pulitzer Prize-winning piece "The Emperor of All Maladies: A Biography of Cancer" (Siddhartha, 2010), and many layers of -omic data can be understood by simple sequence read counts as explained in the "High-Throughput Count Data" chapter of "Modern Statistics for Modern Biology" written by Holmes and Huber

(Holmes and Huber, 2018). Therefore, this dissertation first proposes the idea of generating the sequence read counts for each of the -omic layers. The first chapter seeks to evaluate the possibility of extracting the allelic read counts from over 250,000 sequencing experiments with raw reads deposited in the Sequence Read Archive (SRA), which comprises of over 10,000 public sequencing studies. Not only that the reprocessed data found many applications like extracting variants and identifying allelic specific expressions. The re-processed allelic read counts also enable the creation of a variant-based index, which also helps identifying many variants that were not annotated in the original studies at the time of submission and could be used in studying genetics. The second chapter also asks if it is possible to incorporate even more data layers like transcriptomic and microbiome into our data re-processing effort and create a web-based analysis platform which can simplify the bioinformatics research process, thus enabling the researchers to go from data-querying to analysis with relatively minimal data cleaning.

With the vast amount of reprocessed data, the next logical question to ponder would be if the machine can read the sequencing experiment meta-data and define relevant comparison classes like a human researcher. However, to group the free-text meta-data annotations together into comparison classes would often require human expertise in biomedical language. Chapter three aims to tackle this challenge by repurposing and leveraging the submitter-based BioSamples annotations which are already in a “weakly labeled format" (Peng et al., 2016), to train a natural language processing model to extract the biomedical named entities from the metadata automatically without expert curation.

## References


Crick, Francis. 1970. “Central Dogma of Molecular Biology.” Nature 227 (August). Nature Publishing Group: 561.

Holmes, Susan, and Wolfgang Huber. 2018. “Modern Statistics for Modern Biology.” October 14. http://web.stanford.edu/class/bios221/book/Chap-CountData.html.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” Genome Biology 15 (12): 550.

Peng, Yifan, Chih-Hsuan Wei, and Zhiyong Lu. 2016. “Improving Chemical Disease Relation Extraction with Rich Features and Weakly Labeled Data.” Journal of Cheminformatics 8 (October): 53.

Siddhartha, Mukherjee. 2010. “The Emperor of All Maladies: A Biography of Cancer.” New York, NY: Scribner.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. ’t Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3 (March). The Author(s): 160018.

Yadav, Satya P. 2007. “The Wholeness in Suffix -Omics, -Omes, and the Word Om.” Journal of Biomolecular Techniques: JBT 18 (5): 277.

# CHAPTER 1: Extracting allelic read counts from 250,000 human sequencing runs in Sequence Read Archive

## 1.1 Abstract

The Sequence Read Archive (SRA) contains over one million publicly available sequencing runs from various studies using a variety of sequencing library strategies. These data inherently contain information about underlying genomic sequence variants which we exploit to extract allelic read counts on an unprecedented scale. We reprocessed over 250,000 human sequencing runs (>1000 TB data worth of raw sequence data) into a single unified dataset of allelic read counts for nearly 300,000 variants of biomedical relevance curated by NCBI dbSNP, where germline variants were detected in a median of 912 sequencing runs, and somatic variants were detected in a median of 4,876 sequencing runs, suggesting that this dataset facilitates identification of sequencing runs that harbor variants of interest. Allelic read counts obtained using a targeted alignment were very similar to read counts obtained from whole-genome alignment. Analyzing allelic read count data for matched DNA and RNA samples from tumors, we find that RNA-seq can also recover variants identified by Whole Exome Sequencing (WXS), suggesting that reprocessed allelic read counts can support variant detection across different library strategies in SRA. This study provides a rich database of known human variants across SRA samples that can support future meta-analyses of human sequence variation.

## 1.2 Introduction

The reduction of sequencing cost in recent years (Wetterstrand, 2013) has allowed researchers to progress from sequencing and analyzing a single reference human genome to studying the individual genomes of thousands of subjects (The 1000 Genomes Project Consortium, 2015). The large number of sequencing studies being conducted, together with journal publication requirements for authors to deposit raw sequencing runs in a centralized and open access sequencing archive like Sequence Read Archive (SRA) (Leinonen et al., 2011) have made it possible to perform large scale data analysis on the millions of publically-available sequencing runs.

The SRA contains raw sequencing runs from a variety of projects from large scale consortium studies including Epigenome Roadmap (Roadmap Epigenomics Consortium et al., 2015), ENCODE (ENCODE Project Consortium, 2012), The 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), to small studies being conducted by various independent laboratories. However, the publicly available raw sequencing data are large in size which translates into high storage and computational requirements that hinder access for the broader research community. These requirements can be somewhat mitigated by using preprocessed data such as gene expression matrices, ChIP-seq peak files, or summarized variant information, as such files are much smaller in size. For example, the 1000 Genomes project, The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013) and Genotype-Tissue Expression project (GTEx) (Carithers and Moore, 2015) all offer summarized variant information extracted from the raw sequences in Variant Call Format (VCF) files, containing allelic read counts for both reference and alternative alleles and base quality information which could be used for variant calling.

There have been many efforts to reprocess raw sequencing reads to a more tractable form. However, many of the SRA data reprocessing efforts (Collado-Torres et al., 2017; Lachmann et al., 2017) have focused on quantifying gene expression using public RNA-seq data deposited in the SRA. Sequencing data also capture information about sequence variants, raising the possibility of studying patterns of genetic variation using the SRA. The possibility of extracting variants from RNA-seq was demonstrated on a small scale in a 2015 study (Deelen et al., 2015) where the authors extracted variants using the GATK RNA-seq variant calling pipeline on 5,499 RNA-seq runs in the SRA.

Variant calling typically requires multiple user-specified parameters such as a minimum cut-off for total or read-specific coverage, and usually attempts to model sequencing error explicitly. The primary information used in variant detection is the allelic fraction, the proportion of sequencing reads that support the variant position. Read mapping is highly concordant between alignment tools like bowtie (Langmead and Salzberg, 2012), bwa (Li and Durbin, 2009), novoalign(Ruffalo et al., 2011), supporting the idea, at least for DNA and RNA sequencing experiments, estimates of allelic fraction should be fairly consistent regardless of the specific alignment tool. Using a conservative set of known genetic variants that are unlikely to be the result of sequencing errors, simple filters on coverage or allelic fraction should be sufficient to control error rates at acceptable levels. This would make it possible to collect and analyze known variants across the SRA without applying more complex variant callers.

To explore this possibility, we constructed an allelic read count extraction pipeline to systematically reprocess all available sequencing runs from the SRA. We first applied standard quality filtering to the unaligned reads (see Methods) and then aligned the reads to a subset of the human reference genome that covers 390,000 selected somatic and germline variants curated by the NCBI dbSNP (Kitts and Sherry, 2011) using bowtie2 (Langmead and Salzberg, 2012). To

show that this targeted reference does not introduce unwanted biases into the alignment step, we validated our pipeline performance against alignments performed using whole reference genomes. We next used the TCGA sample-matched Whole Exome Sequencing (WXS) and RNA-seq cohort to confirm that allelic read counts derived from RNA-seq accurately recover variants detected by WXS. We then applied this pipeline to systematically extract variants from over 250,000 sequencing runs in the SRA. Finally, we demonstrated that this allelic read count resource can be used to investigate variants in RNA sequencing studies, even at the single cell level.

## 1.3 Results

***Building a fast allelic fraction extraction pipeline for the SRA.*** As of the end of 2017 the SRA included data from 10,642 human sequencing studies consisting of 697,366 publicly available sequencing runs, encompassing various library strategies such as RNA-seq, WXS, whole genome sequencing (WGS), and ChIP-seq (Methods) and this number continues to increase at a rapid pace (**Figure 1.1**). All of the human sequences deposited in the database were derived from samples carrying germline and somatic variants from the corresponding biospecimen regardless of the original study designs. This presents the opportunity to perform meta-analysis of human genetic variation across studies in the SRA.

However, the complete SRA spans over 1,835 trillion bases, introducing both computational and storage resource requirements that would hinder most researchers from conducting a meta-analysis across many sequencing studies. Therefore, to enable efficient secondary analysis for researchers with limited access to high performance computing (HPC) infrastructure, we sought to process this vast amount of data into a form that can fit on a 1 TB hard disk. To accomplish this, we developed an efficient data processing pipeline (**Figure 1.2**).

We first created a targeted alignment reference that focuses on regions that harbor known variants (n=393,242) curated by NCBI dbSNP (Kitts and Sherry, 2011). These consist predominantly of variants with PubMed references or that have been referenced in selected variant databases (OMIM, LSDB, TPA, or in NCBI curated as diagnostic related). The variants consist mostly of missense mutations with synonymous and truncating mutations accounting for about 15% of the database. Most are germline variants, although the dataset includes a small set of curated somatic mutations. The characteristics of the variants are summarized in **Table 1.1**.

We created the reference alignment index by masking the reference to exclude DNA sequences outside of a region spanning the 1000 base pairs upstream and 1000 base pairs downstream of each variant. This filtering method had been first adopted by Deng et al. to optimize sequencing data processing turnaround times (Deng et al., 2009).

***Large scale allelic read count extraction of human sequence data.*** We retained only sequencing runs from the top five library strategies (RNA-seq, WGS, WXS, AMPLICON, ChIP-seq), and sequencing runs with more than 150 million bases sequenced (equivalent to at least three million reads if the samples have 50 bp per read), corresponding to a total of 304,939 sequencing runs. Of these, 253,005 were successfully processed (**Figure 1.3**) without error with 300 cpu-cores in 30 days. Library strategies were divided between paired-end (64.8%) and single-end (35.2%) sequencing. The difference between the number of pair-end sequencing and single-end sequencing reflects the differing needs of various experimental designs (**Table S1.1**). For example, paired-end sequencing greatly improves the identification of splice isoforms in RNA-seq and structural variants in exome-seq, whereas it provides fewer benefits for other library types that would justify the increased cost relative to single-end sequencing.

One utility that emerges from reprocessing the sequencing data is for imputing experimental annotations. For example, the SRA metadata is not standardized to contain important experimental variables like read length or adaptor sequences, however this information can be easily determined from the raw sequences. A median read length of 95 bp was observed. Most runs (206,360 = 81.56%) had adaptors automatically detected and removed. Sequence and mapping statistics are detailed in the **Table S1.1**. Over these sequencing runs, a median of 2.98% of base pairs were identified as adaptors and were removed. A median base quality Phred score of 36 was observed, suggesting a high overall quality of the sequenced bases in the SRA.

Overall, a median of 296.3 million bases and 10,044,529 read fragments per sample were observed. A median of 5.83% of the reads were aligned to the targeted variant regions (Methods). Adding read length, adaptor contents, number of reads and percentage aligned to the metadata allows the user to better understand the quality of the sequencing runs and filter them accordingly.

***Pipeline performance for targeted variant detection.*** To assess the accuracy of allelic read counts extracted from this targeted reference we compared counts obtained through our pipeline to those extracted from samples pre-aligned to the complete hg38 genome index and downloaded directly from the TCGA. We also took advantage of matched DNA/RNA sequencing in TCGA to evaluate the extent to which allelic read counts extracted from RNA-seq reflect the variants detected from WXS (See section 2.5). We used 524 whole exome tumor sequences from the TCGA Low Grade Glioma (LGG) dataset to assess the performance of our pipeline, as this dataset included the well-known variant (IDH1 R132H) which could serve as a positive control.

The reads from each tumor were aligned to the targeted SNP index and the allelic read counts were compared to the pre-generated alignments available from the TCGA. The resulting variant-locus-by-nucleotide read count matrix contains the read count for each of the four nucleotides across the 393,242 targeted variants at 387,950 genomic sites. We then flattened the nucleic base read count matrix into a single allelic read count vector. For each sample, we compared allelic read counts for all variants obtained using alignment to a targeted reference against allelic read counts obtained from the existing TCGA alignments to a complete reference. Read counts were highly correlated. **Figure 1.4A** shows an example from a single TCGA tumor (UUID: 2b0048e0-a062-40d2-a1e1-4bb763ea0ead), in which a median of 98.2% variants differed less than one log2 fold change in allelic read count from the existing alignment (95% confidence interval: 0.0088 - 0.0554). We found similar correlation across all 524 samples, with a median Pearson correlation (R) of 0.98 for the allelic read counts (95% CI: 0.928 - 0.992; **Figure 1.4B**).

***Effects of PCR duplicates on estimating allelic fraction.*** We next evaluated the necessity of removing putative PCR duplicate reads after alignment based on the extent to which such duplicates bias the estimate of allelic fraction in TCGA. Although most sequence alignment pipelines include a step for removing duplicate reads that result from PCR amplification, recent studies have cast doubt on the benefit of doing so for variant analysis (Ebbert et al., 2016; Stratford et al., 2016). Also, naively removing the duplicated reads could result in overcorrection in high coverage sequencing (Zhou et al., 2014).

We therefore investigated the effect of sequence duplicate removal for all 300k targeted variants across the 524 samples. We compared the allelic read counts extracted with and without duplicate removal for each tumor WXS alignment, and observed a median correlation of 0.983 (95% CI: 0.983-0.990), suggesting duplicate removal had limited impact on allelic read counts. However, we did observe a substantial bias in allelic read count estimates when duplicates are included among sites with very high sequence read coverage. **Figure 1.5A** shows an example using UUID: 0e2c395e-ddda-4833-b1ee-31a9bd08a845. In this sample, deduplicated allelic read counts recover 88.9% of the original allelic read counts among all the variants with ≤100 reads support, while the deduplicated allelic only recover 33.7% of the original allelic read count among all the variants with >100 reads, a 2.63 fold reduction in read count extracted from in the high coverage region (**Figure 1.5A**, slope of grey bar and red bar respectively). Nonetheless, across all 524 samples we observed a difference in allelic fraction < 0.05 for over 90% of the variants when duplicates were excluded, except in extreme cases with over 10,000 mapped reads (median 0.4% of the variants) (**Figure 1.5B**). Thus with high quality sequencing data, filtering duplicates should result in only minor improvement to the data.

***Evaluating variant extraction from RNA-seq using matched DNA/RNA samples.*** The SRA includes over 100k RNA-seq runs and these data contain information about the variant status

of the transcribed DNA. To determine the extent to which variants can be extracted from RNA-seq by our pipeline, we first compared allelic fractions between matched exome sequencing on the one hand and RNA sequencing data in TCGA on the other. TCGA contains samples which have been subjected to both WXS and RNA-seq, which makes it a natural resource for comparing the performance of variant calls derived from RNA-seq data using the WXS-derived variants. We evaluated the possibility of using allelic read counts from RNA-seq to detect both germline and somatic variants.

To evaluate the reliability of allelic read counts for identifying germline variants in RNA sequence reads, we first compared read fractions for germline variants that were homozygous in the corresponding TCGA WXS sample. After collecting all sites that had at least 10 reads and were homozygous for the variant allele in the WXS read data, we evaluated the read counts at those same sites in the RNA-seq data. A median of 5827 sites had at least 10 reads to support the variant in both WXS and RNA-seq for each sample. Across all samples, a median of 97% (95% CI: 95.5% - 97.9%) of sites that were homozygous in the DNA were also found to be homozygous in the matched RNA-seq data.

Next, we explored the utility of allelic read counts for identifying somatic mutations from RNA sequencing data. First, as a positive control, we evaluated the hotspot IDH1 somatic mutation on chromosome 2:208248388 with 395G>A in the template strand, which is most prevalent somatic variant in TCGA LGG on WXS as called by Varscan (Koboldt et al., 2012) (n=371, 70.80% of patients). This variant had been previously identified as enriched in LGG tumors and its status is a major molecular prognostic factor in glioma as noted by the World Health Organization (WHO)(Louis et al., 2016). Using the 524 LGG tumors, we estimated allelic composition using read counts in the matched RNA-seq and WXS independently with our pipeline. The IDH1 mutation status in WXS exhibits a bimodal distribution (**Figure 1.6A**). We selected 10

reads as the cutoff for defining a positive WXS variant. The reference allele was detected in the WXS in all tumors, and 351 patients also had the alternative allele. Over these patients the RNA-seq achieved an area under the precision recall curve (AUPRC) of 0.98 in detecting IDH1 variants observed in the WXS data (**Figure 1.6B**).

We next evaluated the top 100 most frequently observed somatic variants reported by TCGA in the LGG samples that also coincided with the targeted variants, since recurrent mutations are more likely to be drivers and present the most attractive therapeutic targets (Chang et al., 2016). We used the Precision Recall Curve (PRC) framework to determine the extent to which allelic read counts supported expression of the mutant allele. RNA-seq generally recapitulated WXS variants (**Figure 1.6C**), with 70% of the variants having an AUPRC > 0.8, suggesting that majority of the variants called by exome sequencing are expressed in the tumor. However, we do observe 6% of the variants with an AUPRC less than 0.1 when their presence was predicted from RNA-seq allelic fraction. Importantly, these later variants were found in fewer than 10 WXS samples, such that the most recurrent somatic mutations are also more frequently consistently expressed. Thus while absence of a somatic variant cannot be definitively determined from RNA-seq (mutations can be present but not expressed), the most recurrent variants appear to be frequently expressed, suggesting that many somatic mutations of interest will be detectable in RNA-seq data from cancer studies deposited in the SRA.

***Variant landscape of the SRA.*** After validating the general reliability of our allelic fraction estimates, we analyzed 300,000 variants across the SRA. Properties of the variants are listed in **Table 1.1**. Of 300K variants, 170,292 were referenced by PubMed and 138,559 were curated by NCBI as clinically-relevant variants. Out of 156,757 variants with annotated functional effects, the majority were missense mutations (n=91,827). Also, 37,704 variants were annotated as somatic mutations, derived from cancer studies. Overall, the data included a median of three variants per

gene across 21,889 genes. We collected read counts for reference and alternative alleles at these 300K positions for 253,005 human sequencing samples in the SRA. We used default minimum threshold of two reads (Xu, 2018) as the cut-off for Varscan (Koboldt et al., 2012). The distribution of the number of variants are shown in **Figure 1.7**. Known germline variants were detected in a median of 912 sequencing runs, known somatic variants were detected in a median of 4,876 sequencing runs, and known reference alleles were detected in a median of 33,232 sequencing runs. 337 somatic variants, 3,068 germline variants and 23,044 reference alleles were covered by at least two reads in more than half of the sequencing runs, suggesting that SRA data can be repurposed for studying many variants. To facilitate the analysis of variants, we collected allelic read count in each SRA sample into a table (see Data Availability). This read count file allows researchers to rapidly identify which sequencing runs in the SRA have read support for a particular variant.

***Extracting unannotated single cell variants in cancer in SRA.*** Genotype annotations are often missing or incomplete in the SRA, and this limits the reusability of the SRA data. Here, we show that, using the reprocessed data, we were able to recover an important oncogenic mutation BRAF V600E in a single cell RNA-seq study of a patient with myeloid leukemia at diagnosis and as well as at three and six months after diagnosis (Giustacchini et al., 2017). Traditional variant calling relies on high sequencing depths to provide the statistical power to make confident calls. However, since each cell carries only two copies of each chromosome, the low recovery of single cell sequencing makes variant calling from DNA resequencing difficult. Since RNA also contains information about underlying variants and may exist at hundreds of copies per cell (Albayrak et al., 2016), calling variants from single-cell RNA-seq data may circumvent the limitations of DNA resequencing for variants in transcribed regions.

We were able to detect an important oncogenic mutation, BRAF V600E, in single cells using our unified allelic read counts. The overall read depth for the region was 45.9 reads and 17 sites within the 20 bp windows around BRAF V600E had read support for the reference allele. Alternative alleles at the BRAF V600 hotspot were detected in more than 95% cells (**Figure 1.8A**). Also, the alternative allele (T) had a median base quality Phred score of 38 (**Figure 1.8B**) and a median of 22.0 reads to support it (**Figure 1.8C**). Interestingly, we observed a reduction in the reference allele read count over the course of treatment (**Figure 1.8D**) with a corresponding higher fraction of reads supporting the alternate allele, suggesting that the clone with BRAF mutations became more prevalent among the surviving cancer cells, concording with the observation in the study that relapse occurred after treatment.

## 1.4 Discussion

Most published studies on non-protected raw sequencing data are expected to be deposited in the NCBI SRA as a result of journal requirements, and this vast amount of raw sequencing data represents a an opportunity to power large-scale meta-analyses for the interaction of sequence variants with experimental conditions. However, these petabytes worth of sequencing data introduce a computational challenge for analyzing such variants. One solution is to develop a map of relevant sequence variants in the SRA using allelic count profiles.

To create allelic read count profiles from the SRA, we constructed a bioinformatics pipeline with short processing turnaround time by mapping the raw sequencing reads to a targeted reference specific to key somatic and germline variant(s) curated by the NCBI dbSNP. We validated the accuracy of the pipeline by comparing read counts obtained with targeted alignment to counts obtained using complete alignment pipelines, and evaluated genotype consistency across multiple sequencing datasets derived from the same sample. These results confirm that the targeted

alignment pipeline generates allelic read counts that are highly correlated to those from whole genome alignments.

Variant calling has traditionally been performed from DNA sequences, but WXS and WGS library strategies comprise only 40% of the total human SRA data. Thus we also sought to infer the presence of variants from RNA-seq allelic read counts. While RNA may be less reliable for inferring the presence or absence of variants due to gene and allele-specific expression, 61.8% of the RNA-seq samples have more than a million reads mapped onto the targeted variant regions. We also found that highly recurrent somatic mutations detected in WXS of low grade gliomas were also frequently expressed in matched RNA-seq data. Thus, it would also be interesting to utilize the germline allelic read counts extracted from the SRA RNAseq dataset to conduct a large-scale systematic EQTL study. We may also use the somatic allelic read counts in single cell cancer studies to help decipher the interactions between clonal mutations and clonal expressions in tumor heterogeneity.

To the best of our knowledge, this is the first attempt to massively reprocess the human samples in the SRA for the purpose of extracting allelic read counts. The computational infrastructure required to generate variant data at scale presents a barrier to many researchers. Consortia that generate a large volume of sequencing data, such as GTEx, TCGA or the 1000 Genome Project, all offer preprocessed files that enable researchers from the broader community to identify novel findings. Although variant calls are available for some of the datasets included in SRA, significant effort would be required to aggregate these disparate datasets, and most of the non-consortia SRA samples do not have such data available. Simply providing allelic read counts derived through a common bioinformatic pipeline also avoids technical variation that can result from different choice of computational tools and their associated parameter choices. Therefore, we contend that our unfiltered allelic read counts will have broad utility for post hoc analysis.

Many applications require estimates of the magnitude of allelic fraction for inference. This would be particularly useful for questions related to imprinting or reconstruction of tumor subclonal architecture. We found that presence of duplicate reads did not significantly bias estimates of allelic fraction when the quality of the sequencing data is high. However for lower quality datasets or different library strategies, it may still be necessary to remove duplicate reads to obtain high quality estimates. Further analysis is merited to determine which datasets or variants are most confounded if duplicates are not removed. Future releases of the database will include estimates of allelic fraction both before and after removing PCR duplicates.

In conclusion, by reprocessing the raw sequencing runs from the SRA, we improve the findability, accessibility, interoperability, reusability (FAIR) (Wilkinson et al., 2016) of of 250,000 sequencing runs. As the SRA continues to grow, it will be necessary to continuously update the map of variants present in SRA samples. To support variant meta-analyses using the SRA, the next requirement will be unification of the SRA data, including biospecimen and experimental annotations. We anticipate that further refinement of the SRA through efforts such as this will promote reanalysis of existing datasets and lead to new biological discoveries.

## 1.5 Materials and Methods

***SRA Metadata download.*** SRA metadata (files: NCBI_SRA_Metadata_Full_.tar.gz and SRA_Run_Members.tab) were downloaded from ftp.ncbi.nlm.nih.gov/sra/reports/Metadata/ on Jan 4 2018. These files contain the raw freetext biospecimen and experimental annotations. SRA_Run_Members.tab details the relationships between SRA project ID (SRP), sample ID (SRS), experiment ID (SRX) and sequencing run IDs (SRR). We processed only sequencing runs with accession visibility status "public", with availability status "live", and sequencing runs that contains more than 150 million nucleotides bases. We also only included sequencing runs generated from the following library strategies: RNA-Seq, WGS, WXS, ChIP-Seq, AMPLICON. Only samples with layout defined as either SINGLE or PAIRED were considered. We removed SRA study ERP013950 as we noticed it has annotation indicating a total of 85,608 WGS sequencing runs which seem to stem from erroneous submission, as it was only associated with nine biological samples (BioSample) IDs and the experimental annotation was unclear on the nature of the study.

***NCBI dbSNP structure.*** NCBI dbSNP (Kitts and Sherry, 2011) curated a set of SNPs and uses each bit in the bitfield encoding schema to indicate a specific evidence support (ftp://ftp.ncbi.nlm.nih.gov/snp/specs/dbSNP_BitField_latest.pdf). Some evidence supports are derived from databases, for example, NCBI ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), Online Mendelian Inheritance in Man (OMIM, url: https://www.omim.org/), Locus-Specific DataBases (LSDB, url: http://www.hgvs.org/locus-specific-mutation-databases), and Third Party Annotation (TPA, url: https://www.ddbj.nig.ac.jp/ddbj/tpa-e.html). ClinVar contains a curated set of published human variant-phenotype associations. OMIM contains the genotypes and phenotypes of all known mendelian disorders for over 15,000 human genes. LSDB provides gene-centric links to various databases that collect information about variant phenotypes. TPA is a

nucleotide sequence data collection assembled from experimentally determined variants from DDBJ, EMBL-Bank (https://www.ebi.ac.uk/), GenBank, International Nucleotide Sequence Database Collaboration (INSDC) (http://www.insdc.org/), and/ or Trace Archive (https://trace.ncbi.nlm.nih.gov/Traces/home/) with additional feature annotations supported by peer-reviewed experimental or inferential methods.

***Targeted reference building.*** Variants were obtained from dbSNP (downloaded on 4, January on 2017 from ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b150_GRCh38p7/VCF/00-All.vcf.gz), which contained 325,174,853 sites in total, effectively one tenth of our selected human reference genome length (3,099,734,149 bp, version: hg38). We retained only variants with a resource link to any of the existing databases or with support from NCBI curation, indicated by a non zero value for byte 2 of Flag 1 in the NCBI bit field encoding schema, resulting in 393,242 variants. To generate a targeted reference for these variants, we defined 1000 bp downstream and 1000 bp upstream of each SNP as the mapping window. All the regions outside of the windows were masked with base "N" using bedtools v2.26.0 in the reference FASTA file. The reference index was built using bowtie2 v2.2.6(Langmead and Salzberg, 2012) with the merged FASTA file, using default parameters.

***Extracting variants from raw sequencing read FASTQ file.*** We used SRA (Leinonen et al., 2011) prefetch v2.8.0 to download SRR files. Next, fastq-dump v2.4.2 from SRA tool kit was used to extract FASTQ files from SRR into the standard output stream. Trim Galore! version 0.4.0 (url: https://github.com/FelixKrueger/TrimGalore) was then applied to identify adapter sequences using the first 10,000 reads, and the identified adaptor sequence was trimmed in the FASTQ file using cutadapt version 1.16(Martin, 2011), the trimmed reads were then aligned onto the targeted reference (we did not use Trim Galore! to trim the adaptor as it cannot be easily UNIX piped).

Bowtie2 was run with the "--no-unal" parameter to retain only the reads mappable to the target regions in order to minimize the amount of aligned reads for sorting. The alignment file was than sorted using samtools v1.2. and samtools idxstats was used for calculating the number of reads that mapped onto each FASTA reference record. bam-readcount v0.8.0 was used for extracting the per-base allelic read count and per-base quality in the sorted alignment file for each of the targeted genomic coordinates. The paired-end reads were processed the same way as the single-end reads with the exception that paired-end and interleave reads options in fastq-dump, cutadapt, and bowtie2, were specified to ensure proper treatment of paired-end reads. The allelic read counts consist of both the reference allele and alternative allele, and they are retained in the output regardless of the zygosity.

***TCGA download.*** A gdc_manifest was downloaded from the gdc portal on 2017-12-27. We downloaded the TCGA data using gdc-client v1.3.0. We downloaded the associated metadata using the TCGA REST API interface https://api.gdc.cancer.gov/files/. All the alignment files preprocessed from TCGA using GATK pipeline were downloaded. The alignment files were mapped onto GRCh38 with all the raw reads, including read sequence duplicates.

## 1.6 Figures

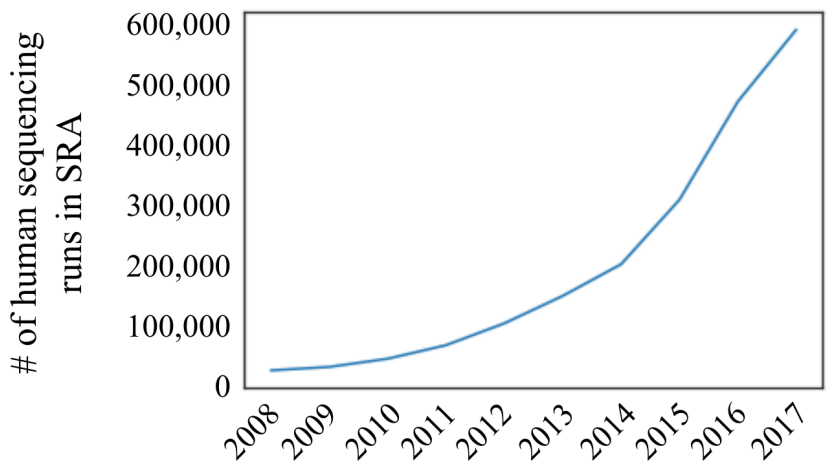



**Figure 1.1. Number of sequencing runs are increasing exponentially in the SRA**

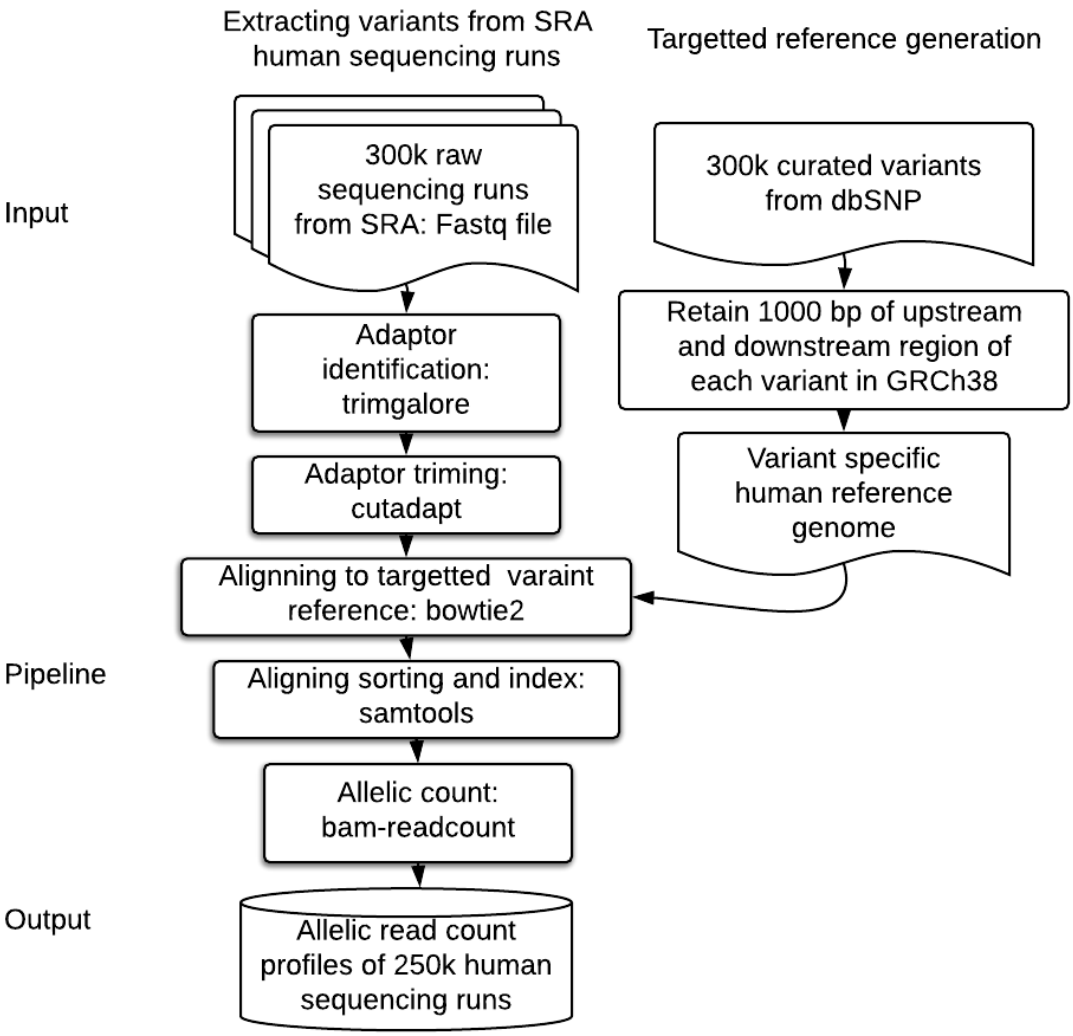



**Figure 1.2. Simple pipeline for extracting >300,000 human sequencing runs from the SRA.** For each sequencing run, first adaptors are identified and trimmed from the raw sequencing reads. Then we align the reads to the targeted reference and extract the allelic read counts.

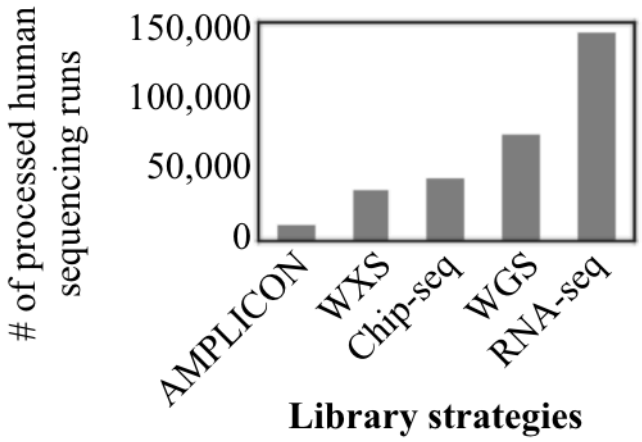



**Figure 1.3. Distribution of the processed SRA data.**

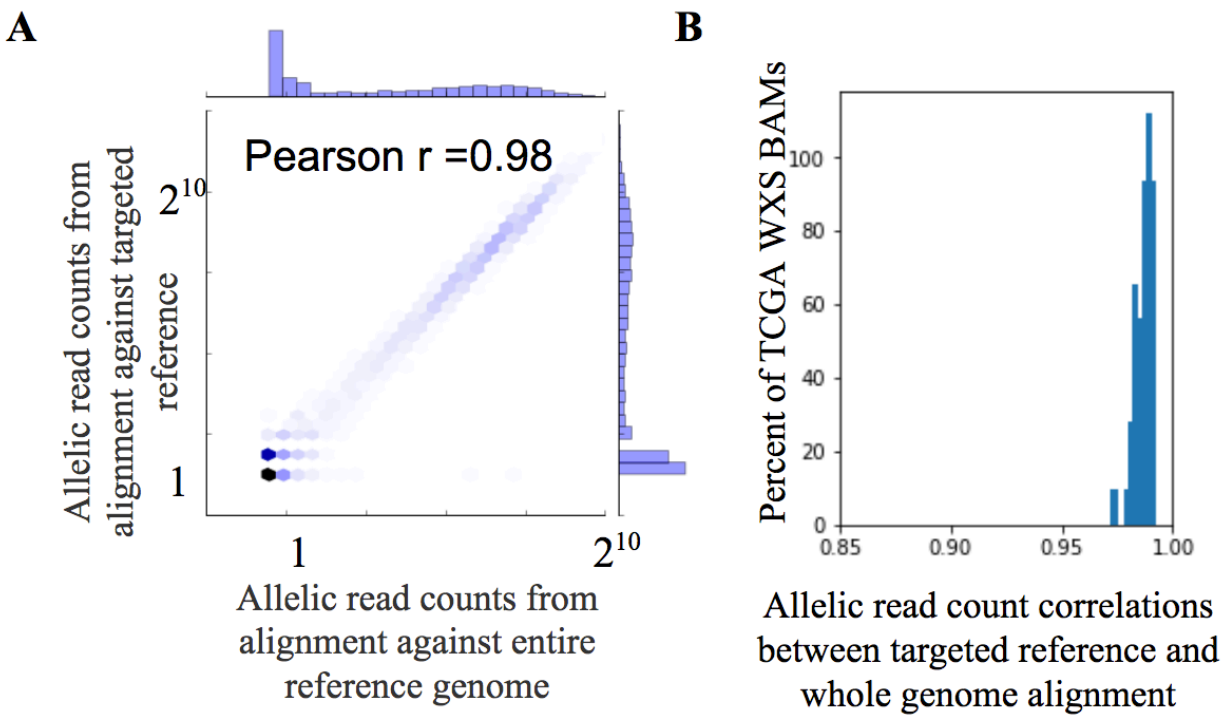



**Figure 1.4. Targeted reference remains accurate for sequence alignment.**
(**A**) Hex density plot showing the high allelic read count correlation between the whole genome alignment (x-axis) and targeted reference alignment (y-axis). Histogram of allelic read counts on whole genome alignment (x-axis, top) and on targeted genome alignment (y-axis, right). (**B**) Distribution of allelic read count correlations (x-axis) over TCGA WXS BAMs (y-axis).

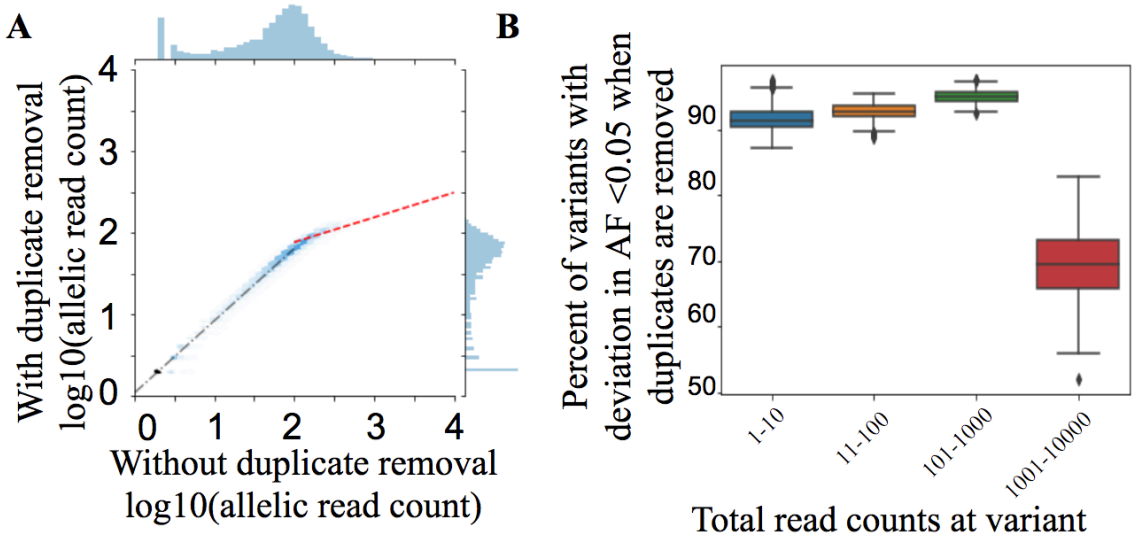



**Figure 1.5. Effect of PCR duplicates removal**
**A** Among regions with <100 reads (grey dashed line), allelic read counts correlate linearly between alignments with duplicate removal (y-axis) and without duplicate removal (x-axis). However, duplicate removal may potentially underestimate read counts in regions with ≥100 reads (red dashed line). **B** Allelic fraction are comparable regardless of duplicate removal except in sites with extremely high read count.

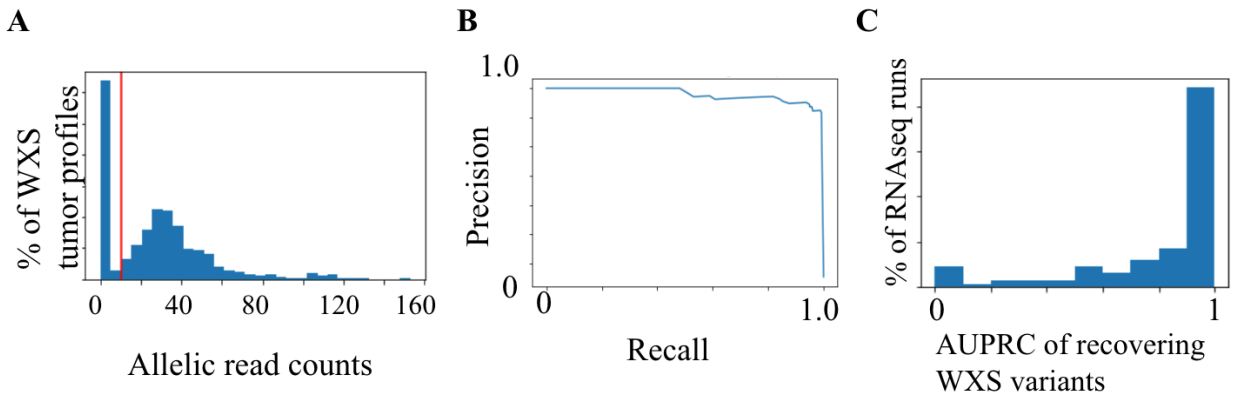



**Figure 1.6. RNA-seq can recover variants extracted from WXS**
(**A**) Minor allelic read counts of IDH1 hotspot mutation. Vertical red line is the binomial distribution cutoff (10 read counts). (**B**) distribution of minor allele of IDH1 (395C>T in template strand). (**C**) RNAseq has high area under the precision recall curve (AUPRC) of recovering WXS variants

## 1.7 Tables

**Table 1.1. Key characteristics of variants in targeted reference**

| Property | Variant type | Number of variants | % of variants |
|---|---|---|---|
| | All | 393,242.00 | |
| | Has 3D structure. SNP3D table | 20,800.00 | 5.29 |
| Resource link property | Cited by PMC article | 170,292.00 | 43.30 |
| | Cited in PubMed or referenced in a clinical database | 201,900.00 | 51.34 |
| Substitution type | Non-synonymous missense | 91,827.00 | 23.35 |
| | synonymous | 32,778.00 | 8.34 |
| | Non-synonymous frameshift | 17,824.00 | 4.53 |
| | Nonsense mutation | 9,286.00 | 2.36 |
| Genotype properties | Genotypes available, also on high density Genotyping kit and have phenotype associations present in dbGaP | 148,114.00 | 37.66 |
| Phenotype properties | Submitted from a locus-specific database | 141,029.00 | 35.86 |
| | Has OMIM/OMIA | 59,617.00 | 15.16 |
| | Somatic (not germline) variant | 37,704.00 | 9.59 |

## Table S1.1. Characteristics and mapping statistics of the SRA sequencing runs.

| | Read statistic | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **AMPLICON** | | **ChIP-seq** | | **RNA-seq** | | **WGS** | | **WXS** | |
| **Library Layout** | **PAIRED** | **SINGLE** | **PAIRED** | **SINGLE** | **PAIRED** | **SINGLE** | **PAIRED** | **SINGLE** | **PAIRED** | **SINGLE** |
| # of sequencing runs | 9472 | 2011 | 6291 | 29410 | 73428 | 48106 | 72915 | 7593 | 26949 | 1497 |
| Median reads per sequencing run | 1704074 | 4024483 | 48970702 | 22147227 | 19204464 | 12001735 | 5677258 | 6544737 | 3849950 | 55985120 |
| Median read length | 250 | 144 | 100 | 50 | 90 | 51 | 101 | 97 | 100 | 179 |
| **Adapter sequences detected by FASTQC** | | | | | | | | | | |
| Illumina TruSeq | 1065 | 904 | 1819 | 12901 | 14993 | 20620 | 16216 | 4379 | 3257 | 1136 |
| Inconclusive auto-detection | 1744 | 1403 | 927 | 19581 | 11630 | 21422 | 16028 | 8598 | 4245 | 2724 |
| Illumina small RNA adapter | 34 | 73 | 1 | 56 | 57 | 1436 | 54 | 519 | 5 | 201 |
| Nextera Transposase sequence | 370 | 213 | 28 | 229 | 3105 | 12053 | 21165 | 3414 | 127 | 64 |
| | **Mapping statistics (median)** | | | | | | | | | |
| Uniquely aligned | 32274 | 199932 | 821671 | 1690868 | 1189300 | 2765875 | 63779 | 383233 | 134891 | 122342 |
| % aligned uniquely in targeted reference | 2.75% | 4.48% | 1.85% | 7.65% | 7.09% | 28.82% | 1.77% | 6.39% | 3.91% | 0.22% |
| Multi-aligned | 4764 | 34196 | 25433 | 104665 | 160681 | 695446 | 2739 | 14256 | 14204 | 11352 |
| % multi-mapped in targeted reference | 0.19% | 0.75% | 0.07% | 0.46% | 0.87% | 6.05% | 0.05% | 0.22% | 0.40% | 0.02% |

### 1.8 Supplementary code and data availability

The python scripts for the pipeline and the jupyter-notebooks for generating the figures are deposited on github (https://github.com/brianyiktaktsui/Skymap) and the data is publicly available on synapse (https://www.synapse.org/#!Synapse:syn11415602). Supplementary table 1 is available on http://hannahcarterlab.org/skymapvariantpsbsupplementarytable1/.

## 1.9 Acknowledgements

We thank all members of the Carter, Mesirov and Ideker lab for scientific feedback and comments. The results here are partly based upon the data generated by the TCGA Research Network: http://cancergenome.nih.gov/. This work was funded by NIH grants DP5-OD017937, RO1 CA220009 and a CIFAR fellowship to H.C.; National Library of Medicine Training Grant T15LM011271 to M.D. Preprint of this article is submitted for consideration in Pacific Symposium on Biocomputing © 2018 copyright World Scientific Publishing Company.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Extracting allelic read counts from 250,000 human sequencing runs" in *Pacific Symposium on Biocomputing*, 2019 by Brian Tsui, Michelle Dow, Dylan Skola, Hannah Carter. The dissertation author was the primary investigator and author of this paper.

## 1.10 References


Albayrak, Cem, Christian A. Jordi, Christoph Zechner, Jing Lin, Colette A. Bichsel, Mustafa Khammash, and Savaş Tay. 2016. "Digital Quantification of Proteins and mRNA in Single Mammalian Cells." Molecular Cell 61 (6): 914–24.

Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." Nature Genetics 45 (10): 1113–20.

Carithers, Latarsha J., and Helen M. Moore. 2015. "The Genotype-Tissue Expression (GTEx) Project." Biopreservation and Biobanking 13 (5): 307–8.

Chang, Matthew T., Saurabh Asthana, Sizhi Paul Gao, Byron H. Lee, Jocelyn S. Chapman, Cyriac Kandoth, Jianjiong Gao, Nicholas D. Socci, David B. Solit, Adam B. Olshen, Nikolaus Schultz, and Barry S. Taylor. 2016. "Identifying Recurrent Mutations in Cancer Reveals Widespread Lineage Diversity and Mutational Specificity." Nature Biotechnology 34 (2): 155–63.

Classes of Genetic Variation Included in dbSNP. 2005. National Center for Biotechnology Information (US).

Collado-Torres, Leonardo, Abhinav Nellore, Kai Kammers, Shannon E. Ellis, Margaret A. Taub, Kasper D. Hansen, Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. 2017. "Reproducible RNA-Seq Analysis Using recount2." Nature Biotechnology 35 (4): 319–21.

Deelen, Patrick, Daria V. Zhernakova, Mark de Haan, Marijke van der Sijde, Marc Jan Bonder, Juha Karjalainen, K. Joeri van der Velde, Kristin M. Abbott, Jingyuan Fu, Cisca Wijmenga, Richard J. Sinke, Morris A. Swertz, and Lude Franke. 2015. "Calling Genotypes from Public RNA-Sequencing Data Enables Identification of Genetic Variants That Affect Gene-Expression Levels." Genome Medicine 7 (1): 30.

Deng, Jie, Robert Shoemaker, Bin Xie, Athurva Gore, Emily M. LeProust, Jessica Antosiewicz-Bourget, Dieter Egli, Nimet Maherali, In-Hyun Park, Junying Yu, George Q. Daley, Kevin Eggan, Konrad Hochedlinger, James Thomson, Wei Wang, Yuan Gao, and Kun Zhang. 2009. "Targeted Bisulfite Sequencing Reveals Changes in DNA Methylation Associated with Nuclear Reprogramming." Nature Biotechnology 27 (4): 353–60.

Ebbert, Mark T. W., Mark E. Wadsworth, Lyndsay A. Staley, Kaitlyn L. Hoyt, Brandon Pickett, Justin Miller, John Duce, Alzheimer's Disease Neuroimaging Initiative, John S. K. Kauwe, and Perry G. Ridge. 2016. "Evaluating the Necessity of PCR Duplicate Removal from next-Generation Sequencing Data and a Comparison of Approaches." BMC Bioinformatics 17 Suppl 7 (July): 239.

ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." Nature 489 (7414): 57–74.

Giustacchini, Alice, Supat Thongjuea, Nikolaos Barkas, Petter S. Woll, Benjamin J. Povinelli, Christopher A. G. Booth, Paul Sopp, Ruggiero Norfo, Alba Rodriguez-Meira, Neil Ashley, Lauren Jamieson, Paresh Vyas, Kristina Anderson, Åsa Segerstolpe, Hong Qian, Ulla Olsson-Strömberg, Satu Mustjoki, Rickard Sandberg, Sten Eirik W. Jacobsen, and Adam J. Mead. 2017. "Single-Cell Transcriptomics Uncovers Distinct Molecular Signatures of Stem Cells in Chronic Myeloid Leukemia." Nature Medicine 23 (6): 692–702.

Kitts, Adrienne, and Stephen Sherry. 2011. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. National Center for Biotechnology Information (US).

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." Genome Research 22 (3): 568–76.

Lachmann, Alexander, Denis Torre, Alexandra B. Keenan, Kathleen M. Jagodnik, Hyojin J. Lee, Moshe C. Silverstein, Lily Wang, and Avi Ma'ayan. 2017. "Massive Mining of Publicly Available RNA-Seq Data from Human and Mouse." bioRxiv. doi:10.1101/189092.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." Nature Methods 9 (4): 357–59.

Leinonen, Rasko, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. 2011. "The Sequence Read Archive." Nucleic Acids Research 39 (Database issue): D19–21.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." Bioinformatics 25 (14): 1754–60.

Louis, David N., Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. 2016. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary." Acta Neuropathologica 131 (6): 803–20.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." EMBnet.journal 17 (1): 10–12.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, Ginell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal,

Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. 2015. "Integrative Analysis of 111 Reference Human Epigenomes." Nature 518 (7539): 317–30.

Ruffalo, Matthew, Thomas LaFramboise, and Mehmet Koyutürk. 2011. "Comparative Analysis of Algorithms for next-Generation Sequencing Read Alignment." Bioinformatics 27 (20): 2790–96.

Stratford, Jeran, Gunjan Hariani, Jeff Jasper, Chad Brown, Wendell Jones, and Victor J. Weigman. 2016. "Abstract 5276: Impact of Duplicate Removal on Low Frequency NGS Somatic Variant Calling." Cancer Research 76 (14 Supplement). American Association for Cancer Research: 5276–5276.

The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." Nature 526 (September). The Author(s): 68.

Wetterstrand, Kris A. 2013. "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)."

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific Data 3 (March). The Author(s): 160018.

Xu, Chang. 2018. "A Review of Somatic Single Nucleotide Variant Calling Algorithms for next-Generation Sequencing Data." Computational and Structural Biotechnology Journal 16 (February): 15–24.

Zhou, Wanding, Tenghui Chen, Hao Zhao, Agda Karina Eterovic, Funda Meric-Bernstam, Gordon B. Mills, and Ken Chen. 2014. "Bias from Removing Read Duplication in Ultra-Deep Sequencing Experiments." Bioinformatics 30 (8): 1073–80.

# CHAPTER 2: SkyMap JupyterHub: A cloud platform to query and analyze >900,000 re-processed public sequencing experiments from >20,000 studies

## 2.1 Abstract

The vast amount of public sequencing experiments generated from human and various model organisms represents a valuable resource for understanding human genetics. However, in their current form, the sequencing data are difficult to reuse. To make these data more readily accessible, we created an efficient pipeline to systematically extract transcript read counts, allelic read counts and microbial read counts from >900,000 sequencing experiments generated from >20,000 studies deposited in the Sequence Read Archive (SRA). This resource could be useful for studying the transcriptomes, genotypes and microbial compositions of >700,000 biosamples. To further make these data accessible to the community, we created a JupyterHub environment where researchers can retrieve these -omic data layers and conduct simple data analysis.

## 2.2 Introduction

Sequencing cost has reduced dramatically in recent years and accelerated the velocity of sequencing data generated from various research bodies. Currently, over millions of sequencing experiments are deposited in the Sequence Read Archive (SRA). However, the preprocessed -omic data associated with each study is often generated by different bioinformatics pipelines, which could result in different data quantification metrics and data formats, making it difficult to aggregate -omic data between studies. Multiple studies (Collado-Torres et al., 2017; Lachmann et al., 2018) had sought to tackle this challenge in RNA-seq, by reducing the raw sequencing data to analysis ready-format like expression matrix. Our previous study (Tsui et al., 2018a) further extended the data re-processing scope beyond transcriptomic, by generating the allelic read counts at the 390,000 dbSNP sites with literature support from over 250,000 human sequencing experiments, which enabled finding unannotated variants in sequencing studies and discovered new associations with experimental conditions that could have been missed otherwise (Tsui et al., 2018b).

Here we expanded our reprocessed data resource to incorporate more species and common -omic layers in our SkyMap database (**Figure 2.1**). Improving and simplifying data access across species is important as animal model remains one of the primary means of studying human disease and phenotype-genotype associations. We believe that increasing the -omic layers is important for our -omic database comprehensiveness. A large proportion of the SRA sequencing experiments were done using bacteria targeted sequencing, and it has been show that it is often possible to repurpose public sequencing data to sample bacterial reads (Zhang et al., 2015) and viral reads (Tang et al., 2013) regardless of library strategy. Thus, we also mapped the reads simultaneously to the viral genomes and the 16S sequences from various bacteria clades. For transcriptome sequencing experiments in the SRA, we also mapped the sequence reads to the reference

transcriptome for the corresponding species. To address the wide range of questions pertaining to the genetic landscape, microbe composition and gene transcriptions as well as in various biosamples collected under different experimental contexts, our reprocessed datasets offer an unified allelic, microbe and expression read count profiles from human, mouse and zebrafish.

We further created a cloud computing platform for bioinformatics analysis that could benefit the -omic research community, where any researcher can input a free-text query to retrieve the data in seconds, and complete some simple -omic analysis with a few clicks without any programming experience. The other useful aspect of our JupyterHub is its capability of creating and accessing an -omic index, where the users can identify studies by variants, expression levels or microbe presence in our reprocessed data.

## 2.3 Results

***SRA data availability.*** The SRA had a total of 8,197.91 trillion bases worth of sequencing data as of December 2018. The SRA sequencing experiments were collected using diverse experimental strategies and collected from various species (**Figure S2.1A**) and the number of sequencing experiments in the SRA is increasing dramatically (**Figure S2.1B**), suggesting the diversity of the SRA data. Here we sought to build a pipeline to reprocess the top 3 SRA library strategies (RNAseq, Amplicon, and Whole genome sequencing (WGS)) which make up 79.39% of the SRA sequencing experiments. RNAseq is the most predominant library strategy (n=1,043,855) likely due to the recent surge of single-cell RNA-seq studies. In addition, a high fraction of metagenome samples (78.41%, n=592,220) dominated the Amplicon sequencing (n=755,249) which is probably due to the standardization of using 16S targeted sequencing in microbial studies. Thus, to enable simple data querying and improve data consistency for this vast amount of sequencing data, we seek to create a simplified sequence reads reprocessing pipeline to generate a unified -omic database consisting of allelic, transcript and microbial read count profile for each sequencing experiment.

***Creating alignment reference to enable sequence read mapping to allelic read and microbe mapping.*** We first extended our the allelic read counts pipeline described in our previous study (Tsui et al., 2018b), which captured the 393,242 human dbSNP sites with variants that are either cited in PubMed or marked as clinically relevant from dbSNP, to capture also the allelic read counts information at the homologous sites in each species. To generate reference indexes for the other species that enable sequencing reads mapping to the homologous sites, we first identified that 71.5% (n=281,359) of the human dbSNP sites are homologous to mouse reference genome, and 35.0% (n=137,512) of the human SNP are homologous to zebrafish reference genome (STAR Methods). We then extended our reference index to also incorporate microbial sequences to detect

the presence of the viral and bacterial sequences. The viral genomes are small, therefore we used the full viral reference genomes from the National Center of Biotechnology Information (NCBI) in the reference index. The 9,555 viral sequences in SRA have a median length of 7,003 bp and has a wide distribution (5% tile=986 bp, 95% tile=172,737 bp, **Figure S2.2A**). The total size of the viral reference is 19.10 Mbp.

The process of mapping short reads to bacterial genomes has been traditionally difficult due to the total size of all bacterial genomes (Truong et al., 2015). 16S sequences are known to be highly conserved within each bacterial clade which could be useful towards identifying microbial community compositions (Clarridge, 2004). Therefore, we used 16S sequences from Mothur (Schloss et al., 2009) as the proxy for bacteria clade counts (STAR Methods). In total, the 13,212 16S sequences have a median length of 1465 bp (95% ci=1267 bp - 1549 bp, **Figure S2.2B**), totalling 260.60 Mbp. The viral and the 16S sequences were then combined with the reference genomes of each species, for the purpose of mapping sequencing reads to both the microbe and reference species. Our pipeline retains only sequencing reads that mapped uniquely to either the corresponding species reference genomes or the microbe references. Among the SRA transcriptomic sequencing experiments, we also enable our pipeline to also map the sequencing reads to the transcriptome. We generated a reference transcriptome index for each species using the reference transcriptomes from Ensembl for the purpose of expression level quantification (STAR Methods), a similar approach from Lachmann et al (Lachmann et al., 2018).

***Landscape of SRA sequence read quality.*** We then used the reference index and built a pipeline (STAR Methods, **Figure S2.3**) to re-process 5,757,288 sequencing experiments prepared using library strategies including Amplicon, ChIp-seq, RNA-seq, WGS and Whole exome sequencing (WXS), where 66.9% of the processed sequencing experiments are pair-ended and 33.1% of the sequencing experiments are single ended. **Table S2.1A** lists the summary quality

check statistics of the sequencing experiments, and 97.95% of base pairs were retained after removing the bases with poor sequencing quality based or if the bases matching with adapter contents (STAR Methods). Unsurprisingly, targeted sequencing library strategies have the longest median read lengths (Amplicon: 175 bp, WXS: 180 bp). These results suggest that majority of the reads from the SRA have high base qualities and reasonable length, and re-mapping of the reads in the SRA sequencing experiments to the references derived from the section above could be automated.

***Landscape of SRA sequence read mappability.*** Our pipeline extracted the allelic read count profiles from a large volume of SRA sequencing experiments. The genome index built with the references mentioned above has successfully extracted the allelic read counts of 905,658 sequencing runs (**Figure 2.2A**) and the sequencing mappability statistics are listed in **Table S2.1B**, where 55.03% of them are from human, 41.18% are from mouse and 2.98% are from zebrafish, suggesting our reprocessed data might be a useful resource towards studying genetics relationships across species. The reference base has read support on a median of 44,751 human sequencing experiments, where a median of 47 sequencing experiment was detected per variant site. In the potential case of lacking existing human biosamples with a specific allele to support a genotype-phenotype association, the user can potentially look at the homologous site from mouse and zebrafish for more evidence support for the association. Each reference base has a median of 20,900 and 2,237 sequencing experiments detected in mouse and zebrafish respectively, and a median of 27 and 6 sequencing experiments to be associated with each variant in mouse and zebrafish respective. These results suggest the utility of cataloging the allelic read counts extracted from the model disease organisms that may have been missed in the original SRA studies.

Next, we sought to evaluate if the sequencing reads from the SRA can also map to the microbial reference. We also observed a high volume of sequencing reads that are mappable

to the microbial references. When we sampled 10,000 sequencing experiments, we observed a median of 33,359 reads being aligned to the microbial genomes, which suggests the possibility of uncovering the microbiome community from the study. We noticed that the major bacterial clades, i.e. Proteobacteria, Firmicutes and Bacteroides have a high volume of sequencing reads in the sequencing experiments (**Figure S2.4A**). We also observed that the major viral class like Hepatitis-C which is associated with various human diseases is detected in over 3.82% of total microbe sequences (**Figure S2.4B**), suggesting that our diverse microbe counts profiles extracted from >900,000 experiments could be useful towards identifying the microbe composition of each sample. However, it would require further study to evaluate how the microbe composition and the microbe read count of each sample differ by experimental conditions and library strategies.

Our pipeline also extracted a large volume of transcript count profiles from the SRA RNA-seq experiments. Lachmann et al. (Lachmann et al., 2018) showed that reprocessing the RNA-seq data can improve the comparability of the data. For the sake for completeness for the SkyMap -omic data layers, we are also providing the reprocessed RNA-seq data. A total of 1,269,678 sequencing experiments were re-processed (**Figure S2.2B**, STAR Methods) using 775,324 biosamples generated from 23,172 studies for human (n=604,917), mouse (n=611,165) and zebrafish (n=31,006). With a median of 60.2% of the reads aligned to the human reference transcriptome, a median of 62.3% of reads aligned to mouse reference transcriptome, a median of 49.0% of reads aligned to the zebrafish reference transcriptome, where majority of the sequencing reads can be mapped without tailoring to the specific RNA-seq library preparation.

***Reprocessed public data simplify hypothesis testing with JupyterHub.*** Often -omic related hypotheses begin with generating and associating the -omic data with experimental conditions. However, the community undermined the availability of the public sequencing-based -omic data due to the laborious process of data aggregation. Here, we provide a simple web-based query

platform for each -omic layers such that the users can go from data querying in each -omic layer to data visualization in less than < 1 minute.

First, we showcase how to user can query and identify the high-quality studies associated with a particular variant at ease using our platform. A variant location for a particular study is indexed if the study has at least 10 sequencing runs and 10 reads associated with the site. For example, if the user is interested in investigating the mutational landscape of BRAF V600 across the existing public studies and its neighboring genomic sites, he can input the corresponding genomic site chr7:140,753,336 in our platform (**Figure 2.3A**), where our platform will then identify all the sequencing studies with the alternative allelic or reference allelic detected at the query position and also the neighbor variants in the 15bp windows (**Figure 2.3B**), which help confirming the mutational status of the biospecimen collected in the SRA. The user can also retrieve the studies with the query mutations using our allelic read count index (**Figure S2.7**). The user can also query the homologous site in the mouse SRA sequencing experiments for further evidence that the site/gene is relevant to the experimental conditions. For example, 217 sequencing experiments also have alternative allele detected at around BRAF V600 in mouse (Chromosome coordinate: chr6: 39,627,782), which increased the number of experiments with BRAF mutation. This example shows how our JupyterHub enables the -omic researchers to search for a particular variant, retrieve and identify all the allelic read counts from the relevant studies efficiently which could then be used for secondary variant analysis.

On the other hand, transcriptome is another important -omic layer that is useful towards understanding how the expressions of different genes contribute to a phenotype or disease. Collado-Torres et al. and Lachmann et al. have shown the utility of providing reprocessed RNA-seq data into the form of expression matrix (Collado-Torres et al., 2017; Lachmann et al., 2018). Here, we a took step further and incorporated the RNA-seq data into our JupyterHub. For example,

if the user in interested conducting a differential expression analysis between B-Cells and T-Cells, the user can supply a free-text query “B-Cell, T-Lymphocytes” which our platform then retrieved the relevant experiments using a substring match and generate an annotated expression matrix (**Figure 2.4A**). Then the user can click on the analysis Jupyter notebooks to visualize the differential expression analysis results (**Figure 2.4B-D**). It is also noteworthy that the Jupyter notebooks plotting integration with Plotly enable the users to edit the figures and table aesthetics at ease (**Figure S2.5**).

Our large volume of also enabled us to collect and generate complex association on a larger scale longitudinal studies. One particular interest was how the gene expression behaves over time and study. For example, TP53 was known to play an important role in the development and stem cell maintenance (Jain and Barton, 2018), where cells with lower expression are thought to be undergoing terminal differentiation (Schmid et al., 1991). Therefore, understanding the expression pattern of TP53 across tissue and its reproducibility across studies is important towards our understanding in cells development. Towards this goal, we collected 20,633 sequencing experiments spanning 865 studies from 45 tissues from 136 time points (**Figure S2.6**). The developmental map overlaid with expression level (**Figure 2.5**) was able to show that Trp53, a homolog of TP53 in human, has a higher expression level in during early tissue development across multiple studies but reduced expression later on.

Aside from variant and transcriptome, we also sought to measure the microbes from the SRA biosamples. Microbiome and virome analysis often involve analyzing the microbe read counts to identify the microbial community that is associated with certain sample conditions, which in some way is similar to the expression analysis. Therefore, we also extended the framework to enable the user to easily identify the microbiome and virome composition in each biosample. For example, the user can quickly confirm that the SRA “HeLa cells” sequencing

experiments have a higher HPV read counts by querying the SRA experiments by free-text query "HeLa cells" and comparing the returned microbe profiles to the sampled sequencing experiments using our platform (**Figure S2.8**). From the case study of HPV read counts, the HeLa cell show higher number of mapped counts than the randomly sampled control sequences (**Figure S2.8**).

## 2.4 Discussion

Currently, if a researcher wants to reuse the public sequencing deposited in the SRA, they need to have access to high-performance-computing (HPC) infrastructure to reprocess the data into the same format. Here, we first reprocessed over 900,000 sequencing experiments to extract the allelic read counts, transcript counts and microbe counts which improved the interoperability and reusability of the vast amount of sequencing data. We then further offered a cloud-based Jupyter-Hub environment which makes the vast amount of reprocessed data become findable and accessible to the research community. Thus, we greatly improved the FAIR-ness (findable, accessible, interpolable and reusable) (Wilkinson et al., 2016) of the sequencing data generated from over 20,000 studies. Also, our current pipeline is updated bi-weekly, which suggest that users without the bioinformatics expertise can submit their data to the SRA directly (which the process is often part of journal requirements) and our pipeline can reprocess the user-submitted data from the SRA, which then the users can then retrieve commonly conducted analysis for their submitted data and compare against the rest of the data at ease using our JupyterHub platform. On top of that, the user can evaluate -omic association across studies easily.

However, there are couple limitations to our current studies. Our analysis platform is still in large limited on -omic layers and offers only simple -omic analysis. For example, we only offer a simple read count and mapping quality score thresholding in variants extraction, which could lose some sensitivity in detecting variants with low allelic read counts, and could be further improved by incorporating statistical modelings (Cibulskis et al., 2013; Koboldt et al., 2012). It would also be interesting to evaluate the batch effect among all the reprocessed -omic layers. Also, we have yet to conduct a thorough investigation of the extracted microbe read count data. It was also unexpected that the single-ended RNA-seq have the highest fraction of reads uniquely aligned to the dbSNP windows (**Table S2.1B**) and the cause will require further investigation.

Aside from addressing and better understanding the limitations listed above, in the future we are also interested in combining the curation-free NLP mechanism from our previous study to automate the process of generating associations between gene sets and experimental conditions, a common task in knowledge curation used in gene sets and gene ontology creation (Ashburner et al., 2000; Liberzon et al., 2011). On the other hand, we are also interested in further expanding our -omic layers to incorporate genome coverage counting profiles for analysis involved in the ChIp-seq and ATAC-seq experiments, and recovering methylated and unmethylated read counting profiles for bisulfite-sequencing experiments.

## 2.5 Materials and Methods

***Create a model chordate organism allelic and microbe mapping pipeline.*** Here, the bioinformatics pipeline is built based on our previously published pipeline but with modifications to allow for both transcript and microbe quantification (**Figure 2.1**). For each sequencing experiment, the pipeline first downloaded the data using SRA prefetch v2.8.0. First, the sequencing runs all underwent sequencing adapter identification using TrimGalore v0.4.0 using the first 10,000 reads and adapter trimming using Cutadapt v1.16. We chose Bowtie2 (Langmead and Salzberg, 2012) as the aligner as it has been shown to be able to map microbial reads (Truong et al., 2015) and gapped reads (Hwang et al., 2015). Then, the sequencing experiments generated using WXS, Amplicon, RNA-seq were all mapped to respective model organism targeted reference genome index using gapped read aligner Bowtie2 v2.2.6 with default global alignments. The model organisms consist of zebrafish and mouse which are model chordate organisms selected by the NIGMS (URL: https://www.nigms.nih.gov/Education/Pages/modelorg_factsheet.aspx).

We then generated a reference FASTA file for each species that allow mapping sequencing reads to the dbSNP sites in each model organism and microbe reference simultaneously. Towards this goal, we first mapped the homologous sites using the University of California, Santa Cruz (UCSC) LiftOver (version 366) with default parameters and generated an SNP reference specific genome using the approach similar to our previous study (Tsui et al., 2018b). The homologous sites was mapped with the UCSC Chain files that provided the confident alignments between human and mouse at URL: http://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToMm10.over.chain.gz, and with the UCSC Chain files that contain the alignment between human and zebrafish at URL: http://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToDanRer10.over.chain.gz, where the chain files were downloaded at 1, Oct. 2017. We defined 1000 bp downstream and 1000

bp upstream of each SNP as the mapping window to generate a targeted reference for these variants. The non-mappable regions outside of the windows were masked with base "N" using Bedtools v2.26.0 in the chordate species reference FASTA file.

The microbe reference FASTA records also consist of the 16S sequences from Mothur (URL: https://www.mothur.org/wiki/RDP_reference_files) and also the viral reference sequences from the NCBI (URL: ftp://ftp.ncbi.nih.gov/refseq/release/viral/ ). Therefore, for each species we concatenated the FASTA records from the chordate species SNP reference with the microbe reference. The reference alignment index was built using Bowtie2 v2.2.6 with reference genome in the format of FASTA records.

The sequencing reads aligned in the format of BAM were then sorted using Samtools v1.2 and Samtools idxstats was then used in quantifying the number of reads that were mapped onto each reference record. Bam-readcount v0.8.0 was then used for extracting the per-base allelic read counts in the sorted alignment file for each of the targeted genomic coordinates. The outputted allelic read counts consist of both the reference allele and the alternative allele, and they are retained in the output as long as they have >1 read mapping support. We then extracted the number of reads aligned to each microbe FASTA record in the BAM file using Samtools idxstats v1.2.

***Create an RNA-seq pipeline for all the SRA data.*** We also offer a unified RNA-seq resource. The RNA-seq resource building process involves first building a reference transcriptome index with the transcriptome aligner Kallisto (Bray et al., 2016) (V.44.0) with default parameters, using the reference FASTA transcriptome files Mus_musculus.GRCm38.cdna.all.fa.gz from the mouse, human reference with file name Homo_sapiens.GRCh38.cdna.all.fa.gz and zebrafish reference with file name Danio_rerio.GRCz10.cdna.all.fa.gz downloaded from the Ensembl FTP site (URL: ftp.ensembl.org/pub/release-94/fasta/homo_sapiens/cdna/ ). Each RNA-seq sequencing experiment in the format of FASTQ was downloaded using SRA prefetch v2.8.0, and

had their adaptors identified using Trimgalore (Krueger, 2015) and adaptors trimmed using Cutadapt (Martin, 2011). The trimmed sequencing reads were then UNIX piped to Kallisto quantification process with default parameters. The output transcripts were quantified by the normalized metric transcript per million (TPM) and estimated transcriptome counts. The TPM counts profiles and estimated count profiles for each sequencing runs were then merged as a single matrix. The transcripts were then summed and merged into gene-level quantification using the transcript to gene mapping offered within the annotation section of each transcript record in the Ensembl reference FASTA file. All the RNA-seq profiles were then merged into a single python NumPy (URL: http://www.numpy.org/) expression matrix which offers the native memory mapped mechanism to extract the expression level either by gene or by sequencing runs efficiently without the need of preloading the entire data matrix in memory which might incur overhead in loading time or cause memory strain.

***Setup of the JupyterHub***. We leveraged the power of the latest cloud tools to improve service availability and to reduce the maintenance burden. We followed the guide from the official JupyterHub guide to set up the SkyMap JupyterHub (URL: https://zero-to-jupyterhub.readthedocs.io/en/latest/). This setup utilizes the container management service Kubernete to manage the cloud resource, which can automate the process of computational resource scaling and fault tolerance, to ensure the quality of our user experiences. We stored the SkyMap data on the Amazon Web Service (AWS) Elastic File Storage (EFS), which behave like a read-only UNIX file system and scale across many Elastic Instances (EC). The UNIX file system allows us to memory map the expression matrix in NumPy format wrapped in Pandas, which act as a data frame which the user can query the matrix using the native Pandas syntax. Our code primarily relies on such commonly used packages in the data analytics in Python, which can maximize the code understandability. Also, to ensure reproducibility, our Jupyter notebooks,

Conda (URL: https://conda.io/docs/contributing.html) package environment, Python scripts and Skymap Docker image are all open source and available on GitHub repository at URL: https://github.com/brianyiktaktsui/Skymap. We use T2.xlarge instance for head nodes and slave nodes in the cluster. All the nodes are running on a regular EC2 instance with Kubernete image (AWS Amazon Image ID ami-009b9699070ffc46f), with the exception of using the AWS Spot instances on all the slave nodes to reduce the financial cost.

***Data availability.*** We currently reprocess the SRA data bi-weekly and make the data available in the JupyterHub. Our cloud based JupyterHub platform is located at URL: http://jupyterhub.hannahcarterlab.org with login requirement, which in the future we intend to provide a interface for user to upload their own private data to be processed and compared against the public SkyMap data. The re-processed data is available on the FTP site (ftp://download.hannahcarterlab.org/) without login requirement which allow bulk download from the users.

## 2.6 Figures

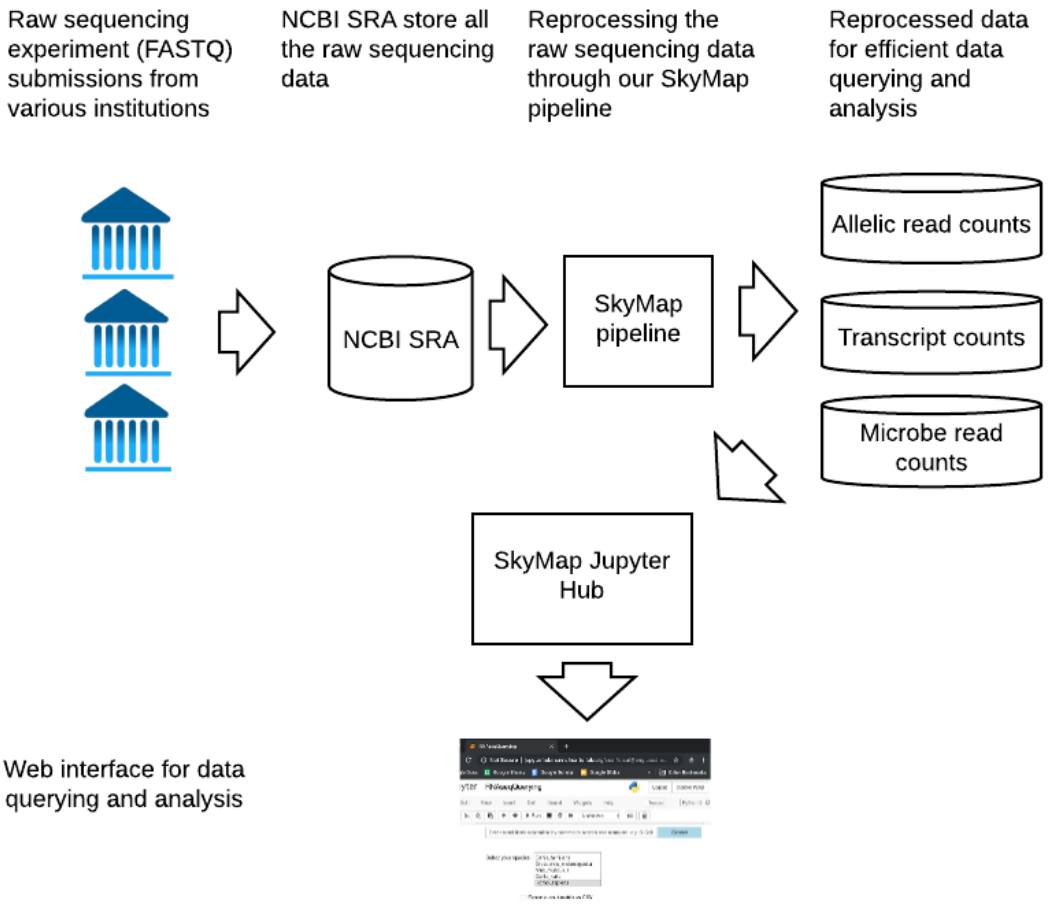



**Figure 2.1. Overview**
Reprocessing the SRA to enable simple data access to the broad -omic community.

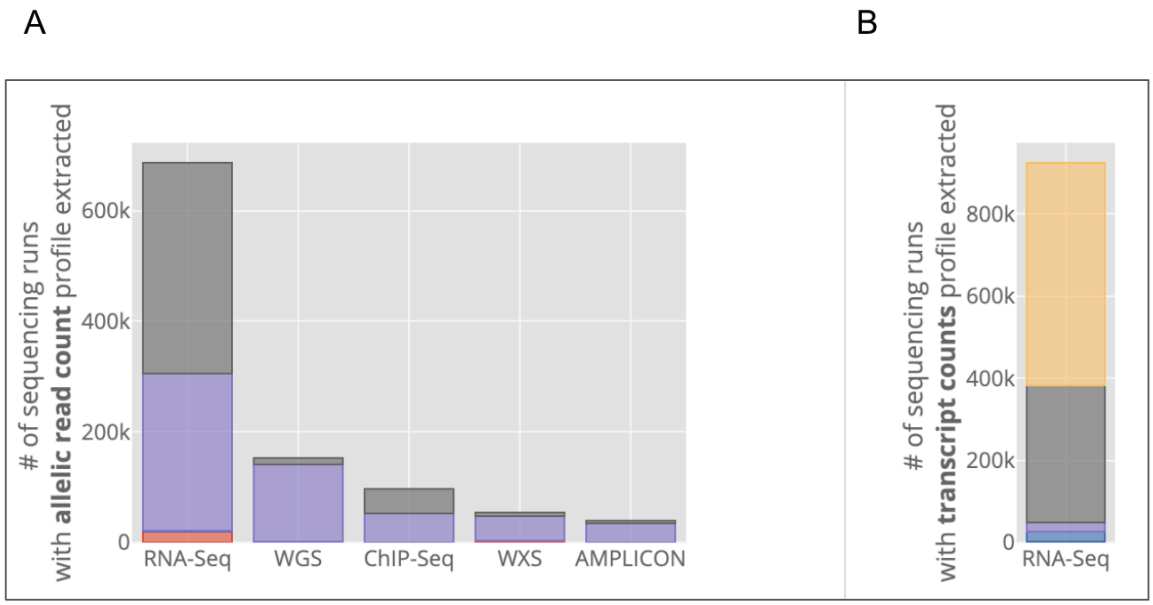
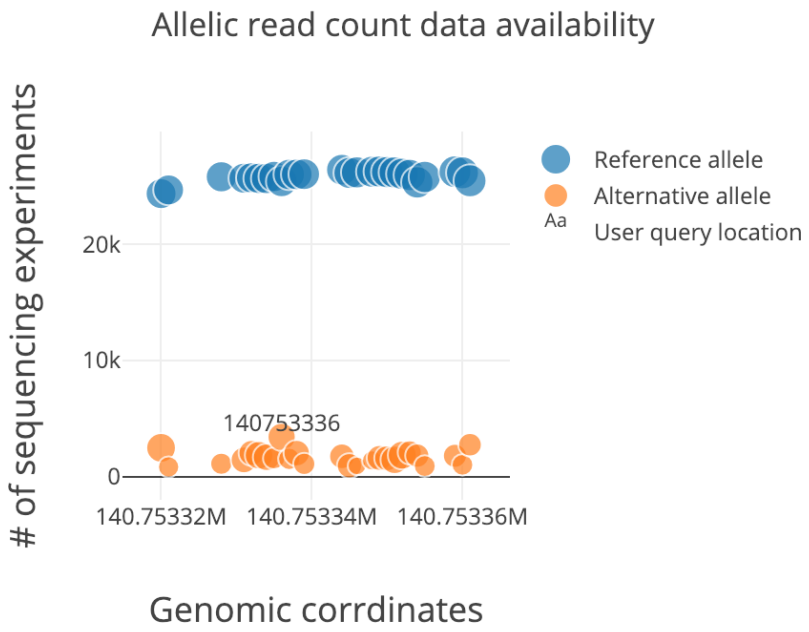



**Figure 2.2. The landscape of reprocessed SRA data**
(A) Data availability of the allelic read counts data extracted from the SRA. (B) data availability of the transcriptome data extracted from the SRA.

A




B

| # of SRR in the study | Study annotations |
|---|---|
| 418 | SRP accession ID: SRP067759<br>Title: b'Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia'<br>Study description: b'Single-cell whole transcriptome analysis of chronic myeloid leukemia stem cells, defined design: We developed a method that combines high-sensitivity mutatio transcriptome analysis of the same single cell. We applied this techniqu single cells from patients with chronic myeloid leukemia (CML) through revealing heterogeneity of CML stem cells, including the identification o cells with a distinct molecular signature that selectively persisted during performed differential gene expression and the gene set enrichment an single cells from either patients or normal donors, analyzed at different |
| 392 | SRP accession ID: ERP014537<br>Title: b'A generic, cost-effective and scalable cell lineage analysis platform'<br>Study description: b'Advances in single-cell genomics enable commensu in methods for uncovering lineage relations among individual cells. Cur for cell lineage analysis depend on low resolution bulk analysis or rely o sequencing which is not scalable and could be biased by functional dep integrated biochemical-computational platform for generic single-cell li retrospective, cost-effective and scalable. It consists of a biochemical-co |

**Figure 2.3. Variant extraction using our JuypterHub.**
If the user is interested in investigating the mutational landscape of BRAF V600 across the existing public studies and its neighboring genomic sites, he can input the corresponding genomic site chr7:140,753,336 in our platform (A) where our platform will then identify all the sequencing studies with the allelic read counts detected at the query position and also the neighbor variants in the 15bp windows, which is informative towards evaluating the studies with the query mutation. (B) Example interactive study summary table display: The table display is the screenshot of our interactive JupyterHub web interface which allows the user to scroll through the available studies ranked by the number of sequencing experiments with alternative allele detected.

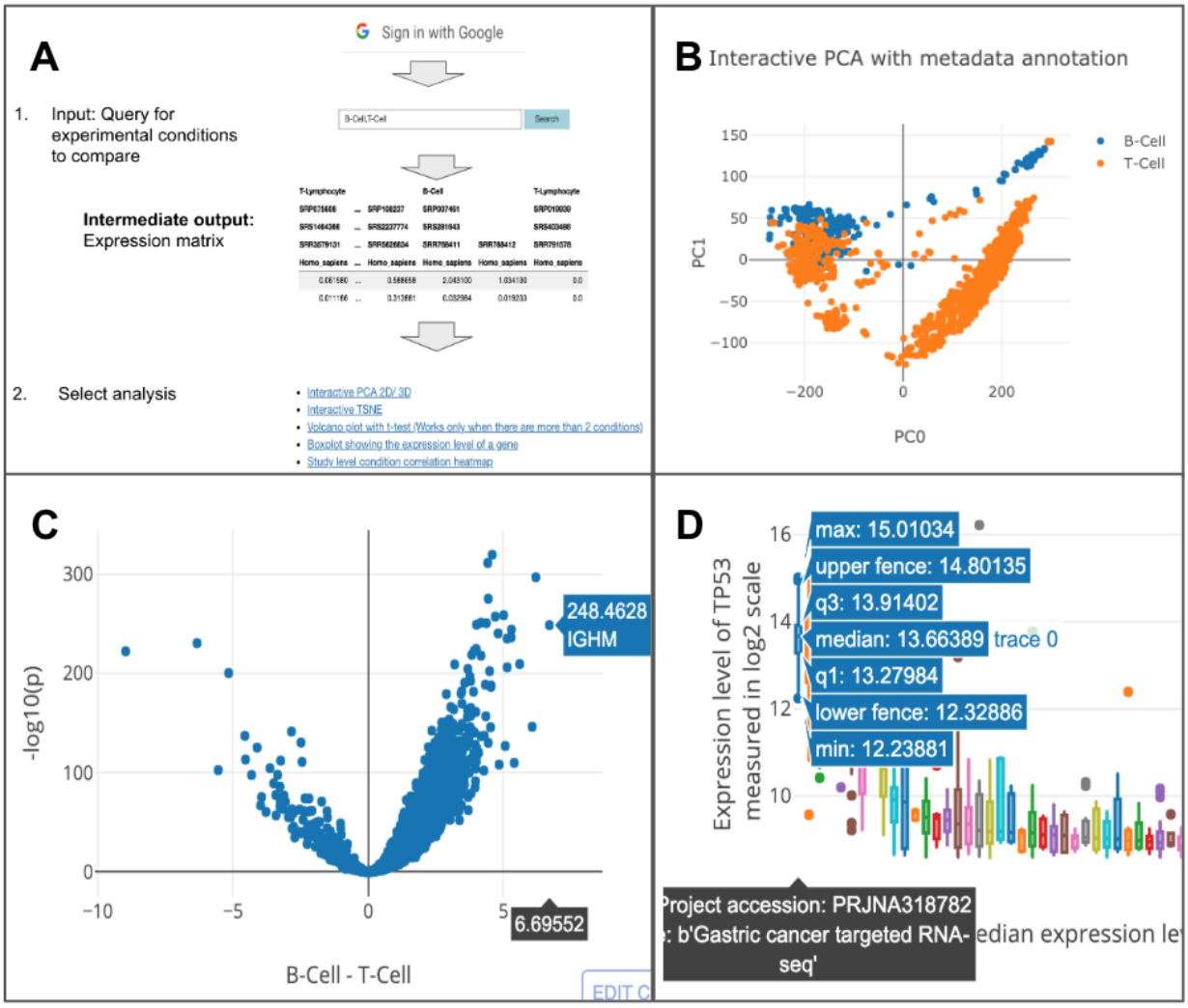



**Figure 2.4. Expression analysis using our JupyterHub.**
This example shows how the user can query and compare the expression profiles between T-Cell and B-Cel using our JupyterHub in less than minutes. (A) (top) Workflow: the user input a list of experimental conditions. In this example, “B-Cell, T-Cell” as a free-text query, (middle) which will retrieve the expression profiles with annotations containing "B-Cell" or "T-Cell'. (bottom) The user could click on the analysis links to generate various differential expression analysis. Example Analysis: (B) PCA showing that the expression profiles are separated by the first two PCs. (C) Interactive volcano plot to identify DE genes. (D) The user can also query by different gene names to retrieve the studies with the highest expression levels as opposed to querying by metadata.

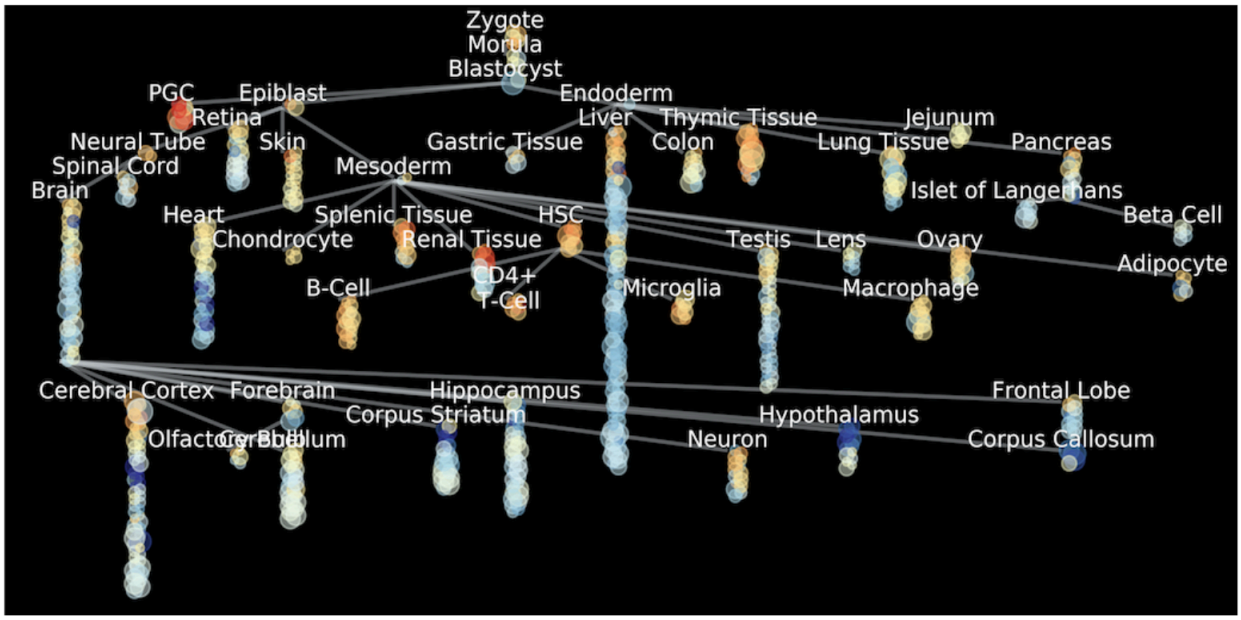



**Figure 2.5. Large-scale meta-analysis using reprocessed RNA-seq data.**
Large-scale meta-analysis of expression levels of Trp53 (high expression level: red, low expression level: blue) across different studies (node). The studies within each tissue were ranked vertically according to the developmental timeline within each tissue (component). Node sizes represent the study size in log2 scale. The reference timeline can be found at Figure S8.

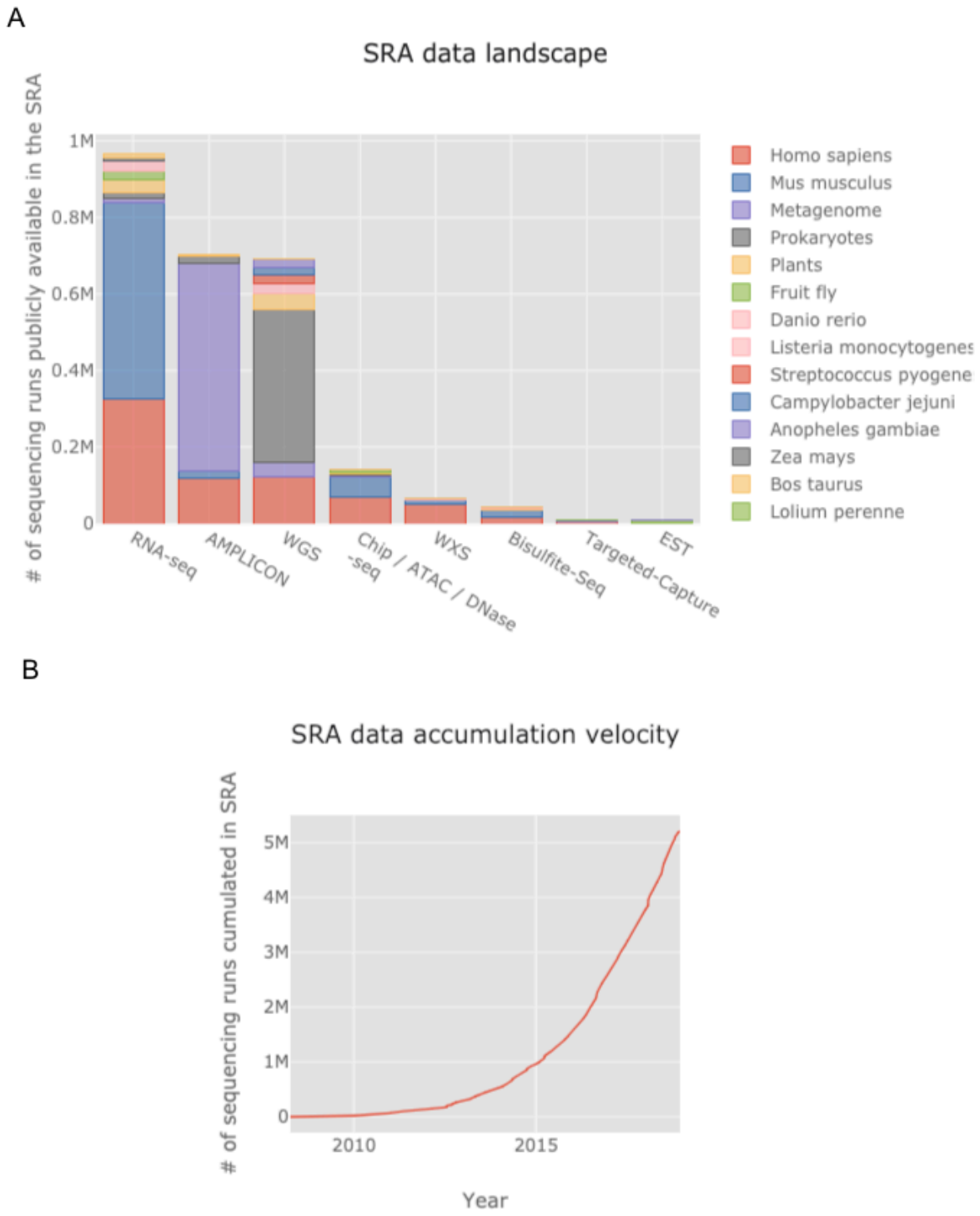



**Figure S2.1: SRA data landscape.**
(A) SRA data volume (y-axis) across the library strategies with highest availability (x-axis) and species with highest availability (hue). (B) SRA data volume (y-axis) is increasing over time (x-axis).

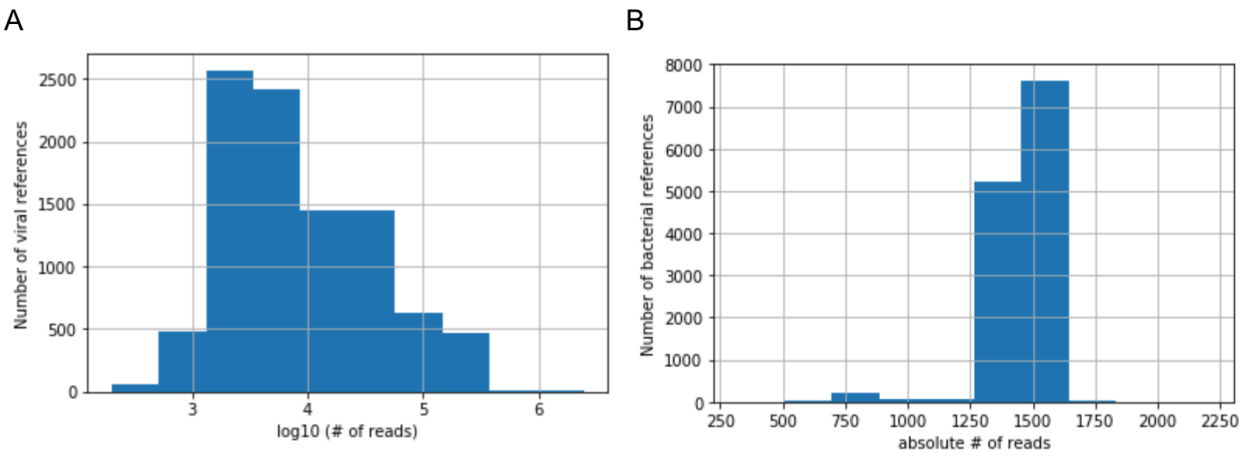



**Figure S2.2: Microbe read lengths.**
(A) histogram of viral sequence length (B) histogram of 16S sequence length.

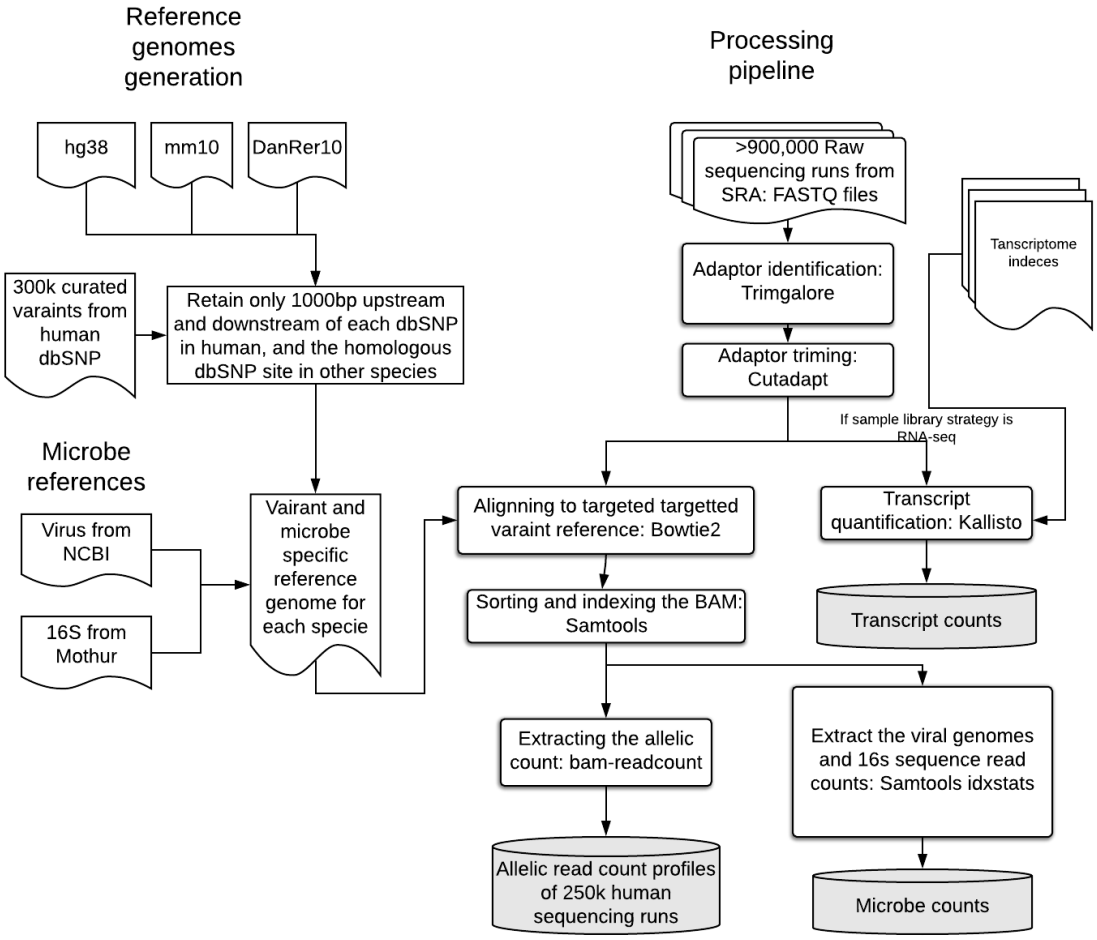



**Figure S2.3: Workflow diagram.**
(Left) The reference species dbSNP index was combined with the 16S sequences and viral genome sequences to form a single reference for each species to align to; (Right) Sequencing pipeline to generate allelic read counts, transcript counts and microbial counts using the generated reference indices.

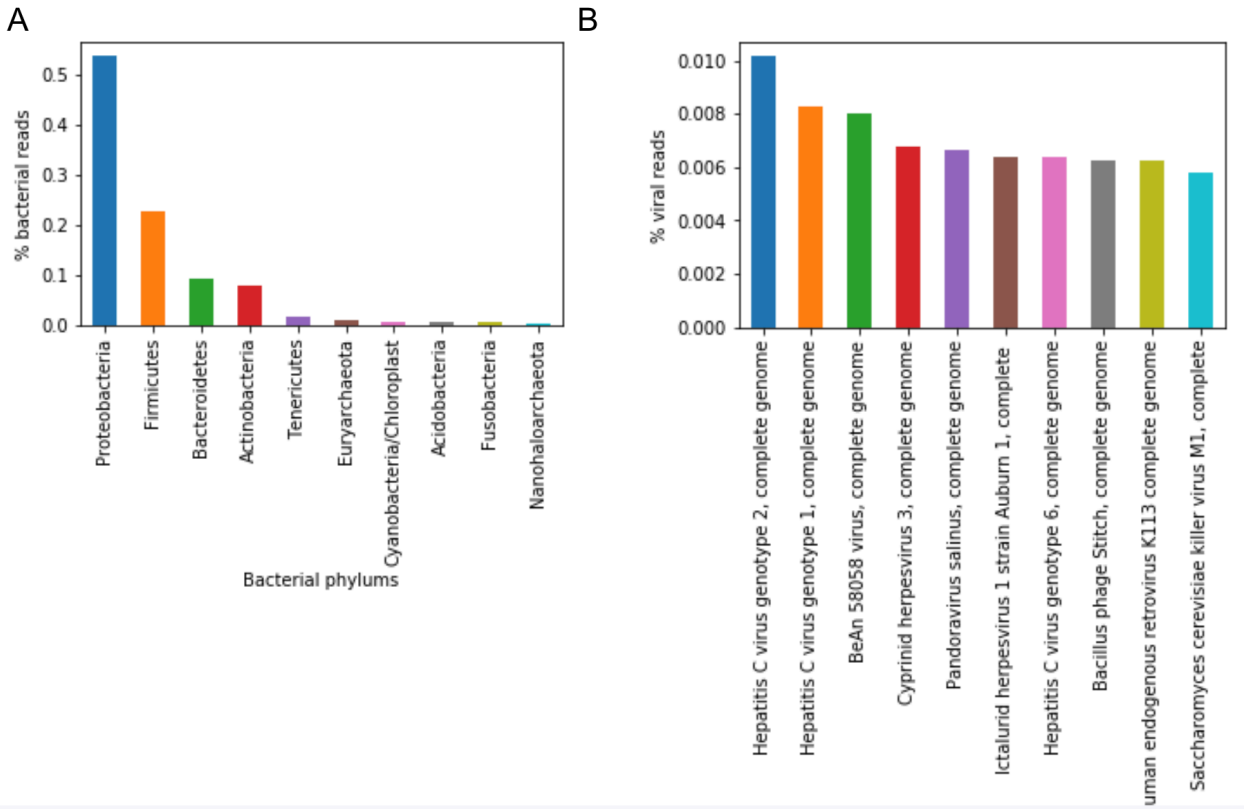



**Figure S2.4: Distribution of microbe abundance in the SRA.**
(A) Bacterial phylums defined by Mothur (y-axis); percentage read abundance in the SRA (x-axis); (B) viral phylum defined by NCBI (y-axis); percentage read abundance in the SRA (y-axis)

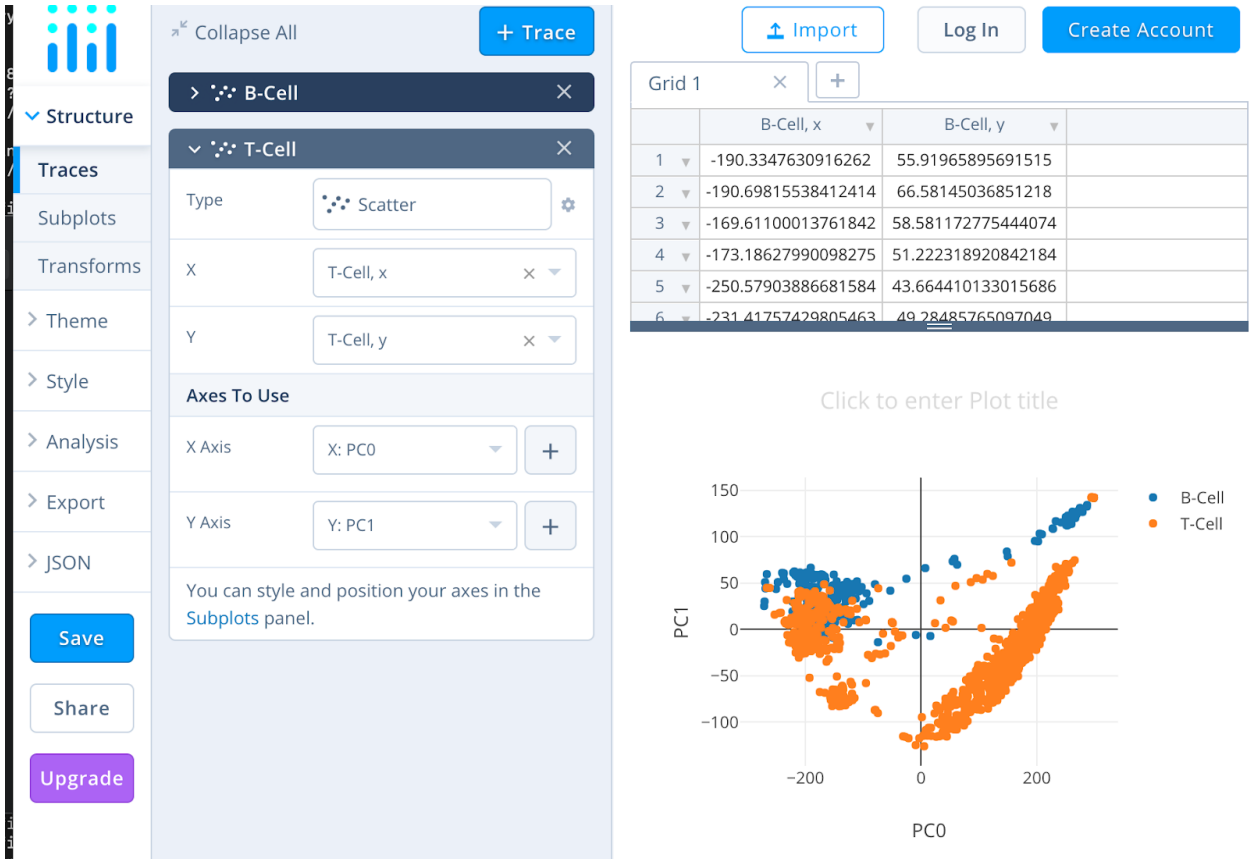



**Figure S2.5: Plotly cloud user interface.**
Our JupyterHub support Plotly plotting which enables the user to edit the figures and table aesthetics after exporting the table from Jupyter notebook to Plotly cloud (URL: https://plot.ly/products/cloud/)

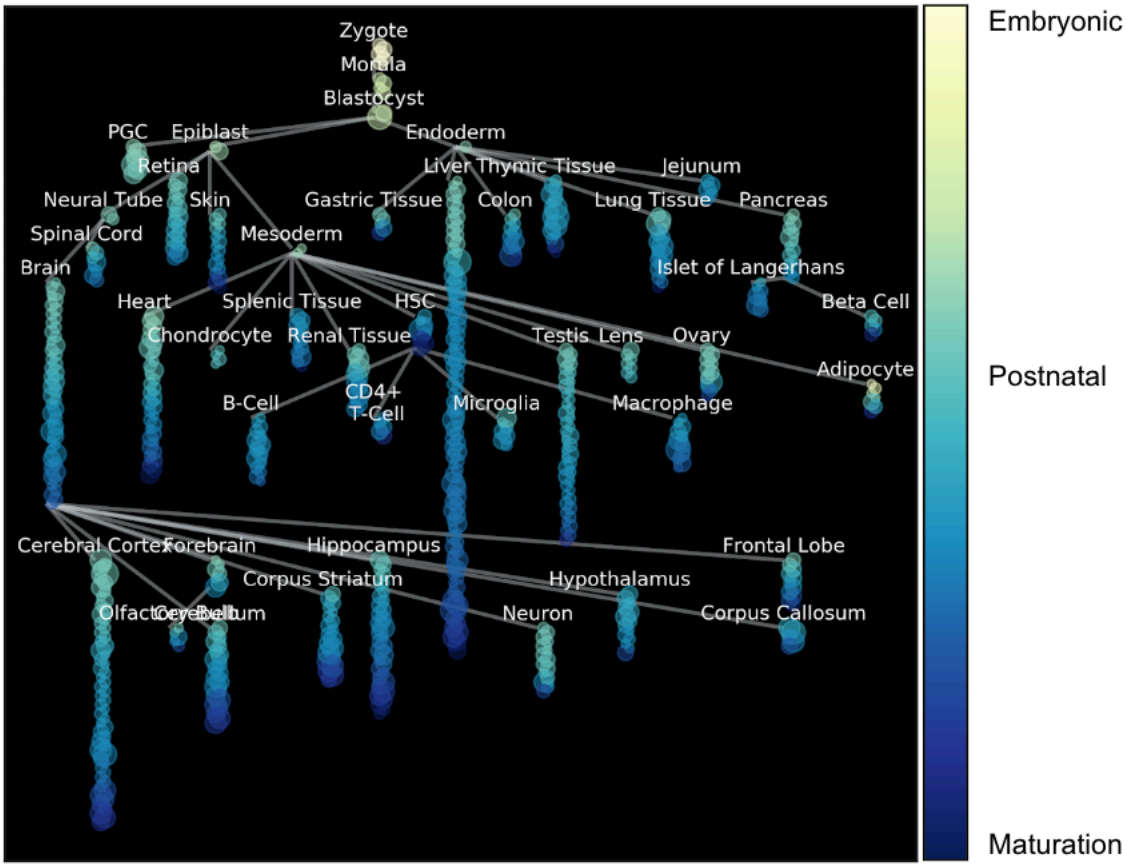



**Figure S2.6: Reference developmental timeline.**
The developmental time are colored according to the developmental timeline extracted from the metadata in log2 scale, where each node in the tissue represent a study with data collected at a particular time point, and an edge represent a differentiation relationship or an anatomic "part_of" relationship.

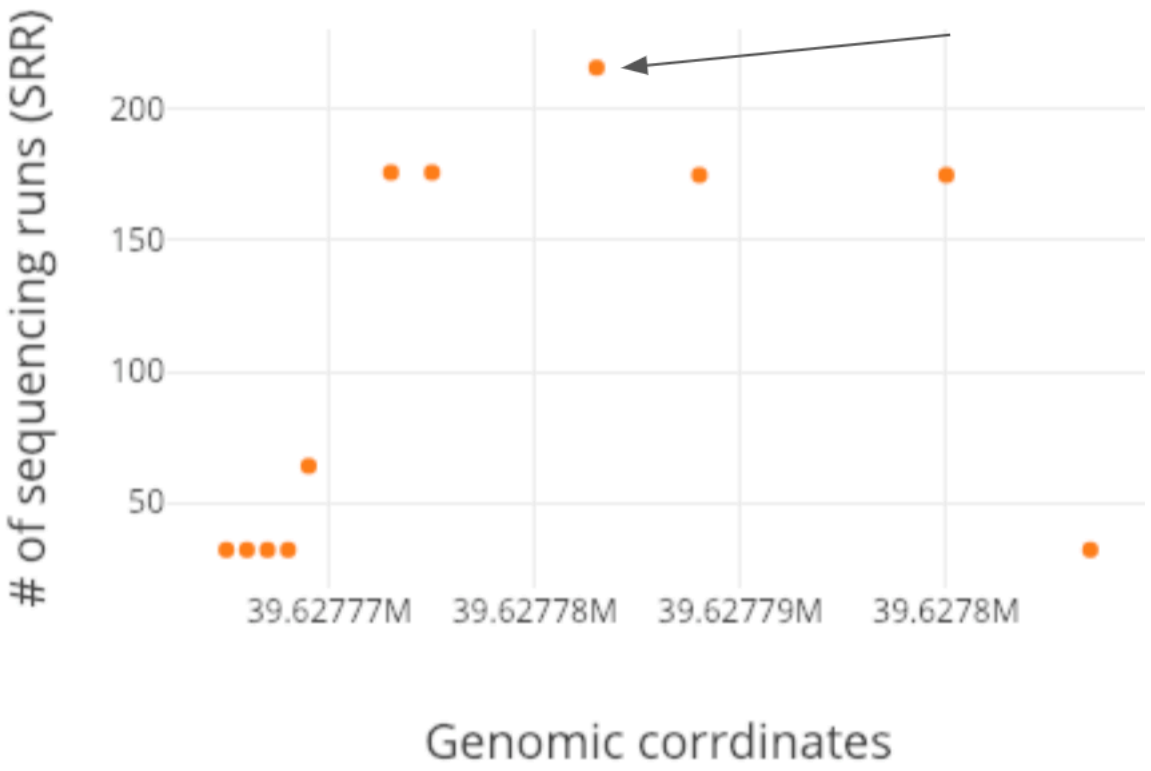



**Figure S2.7: Identifying sequencing experiments in the mouse with human homolog variants.** In this example, the user can query the homologous site of human BRAF V600 in mouse, to find the number of sequencing runs with variation (y-axis) at the homologous site and also the neighboring genomic sites (x-axis).

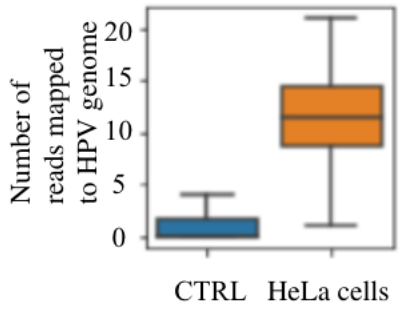



**Figure S2.8: Microbe analysis using SkyMap JupyterHub.**
By querying the SRA experiments using the free-text query “HeLa cells”, the user can compare the HeLa cell microbe profiles to the randomly sampled control (CTRL) sequencing experiments using our platform.

## 2.8 Table

**Table S2.1. Quality check statistics and sequence alignment statistics.**
(A) summarizing the median number of percentage base written, the median read length, and the percentage reads with adapters in the SRA for each library strategy and library layout. (B) summarizing the median number of sequence reads detected in the SRA and the percent of sequence reads uniquely aligned to our custom allelic read and microbe read mapping reference.

| | Library Strategy | AMPLICON | | ChIP-Seq | | RNA-Seq | | WGS | | WXS | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Library Layout | PAIRED | SINGLE | PAIRED | SINGLE | PAIRED | SINGLE | PAIRED | SINGLE | PAIRED | SINGLE |
| **A** | | | | | | | | | | | |
| | % bases written | 95.39% | 92.08% | 96.02% | 96.98% | 87.85% | 98.56% | 96.92% | 66.11% | 99.13% | 94.14% |
| | Read length | 175 | 125 | 75 | 50 | 90 | 51 | 101 | 76 | 75 | 180 |
| QC statistics | % reads with adapters | 6.86% | 6.09% | 7.15% | 8.86% | 7.13% | 8.68% | 6.40% | 8.17% | 6.70% | 8.03% |
| **B** | | | | | | | | | | | |
| | Number of reads detected | 3.03E+06 | 5.67E+06 | 4.05E+07 | 2.71E+07 | 1.22E+07 | 1.69E+07 | 1.75E+07 | 1.04E+07 | 2.36E+07 | 3.56E+07 |
| Sequence alignment statistics | % uniquely aligned to the custom reference | 5.61% | 24.14% | 1.88% | 7.50% | 7.27% | 26.64% | 1.21% | 5.37% | 4.62% | 8.85% |

## 2.10 Acknowledgements

This work was supported by CIFAR fellowship to H.C., and RO1 CA220009 to M.Z. and H.C., P41-GM103504 for the computing resources.

Chapter 2, in part, is a reformatted presentation of the material currently being prepared for submission for publication as "SkyMap JupyterHub: A cloud platform to query and analyze >900,000 re-processed public sequencing experiments from >20,000 studies" by Brian Tsui, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

## 2.11 References


Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." Nature Genetics 25 (1): 25–29.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." Nature Biotechnology 34 (5): 525–27.

Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. 2013. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." Nature Biotechnology 31 (3): 213–19.

Clarridge, Jill E., 3rd. 2004. "Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases." Clinical Microbiology Reviews 17 (4): 840–62, table of contents.

Collado-Torres, Leonardo, Abhinav Nellore, Kai Kammers, Shannon E. Ellis, Margaret A. Taub, Kasper D. Hansen, Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. 2017. "Reproducible RNA-Seq Analysis Using recount2." Nature Biotechnology 35 (4): 319–21.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." Nucleic Acids Research 30 (1). Oxford Univ Press: 207–10.

Hwang, Sohyun, Eiru Kim, Insuk Lee, and Edward M. Marcotte. 2015. "Systematic Comparison of Variant Calling Pipelines Using Gold Standard Personal Exome Variants." Scientific Reports 5 (December): 17875.

Jain, Abhinav K., and Michelle Craig Barton. 2018. "p53: Emerging Roles in Stem Cells, Development and beyond." Development 145 (8). doi:10.1242/dev.158360.

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." Genome Research 22 (3): 568–76.

Krueger, Felix. 2015. "Trim Galore." A Wrapper Tool around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files.

Lachmann, Alexander, Denis Torre, Alexandra B. Keenan, Kathleen M. Jagodnik, Hoyjin J. Lee, Lily Wang, Moshe C. Silverstein, and Avi Ma'ayan. 2018. "Massive Mining of Publicly Available RNA-Seq Data from Human and Mouse." Nature Communications 9 (1): 1366.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” Nature Methods 9 (4): 357–59.

Leinonen, Rasko, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. 2010. “The Sequence Read Archive.” Nucleic Acids Research 39 (suppl_1). Oxford University Press: D19–21.

Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. 2011. “Molecular Signatures Database (MSigDB) 3.0.” Bioinformatics 27 (12): 1739–40.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” EMBnet.journal 17 (1): 10–12.

Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, Brian B. Oakley, Donovan H. Parks, Courtney J. Robinson, Jason W. Sahl, Blaz Stres, Gerhard G. Thallinger, David J. Van Horn, and Carolyn F. Weber. 2009. “Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.” Applied and Environmental Microbiology 75 (23): 7537–41.

Schmid, P., A. Lorenz, H. Hameister, and M. Montenarh. 1991. “Expression of p53 during Mouse Embryogenesis.” Development 113 (3): 857–65.

Tang, Ka-Wei, Babak Alaei-Mahabadi, Tore Samuelsson, Magnus Lindh, and Erik Larsson. 2013. “The Landscape of Viral Expression and Host Gene Fusion and Adaptation in Human Cancer.” Nature Communications 4: 2513.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. “MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling.” Nature Methods 12 (10): 902–3.

Tsui, Brian Y., Michelle Dow, Dylan Skola, and Hannah Carter. 2019. “Extracting Allelic Read Counts from 250,000 Human Sequencing Runs in Sequence Read Archive.” Pacific Symposium on Biocomputing, January, 196–204.

Tsui, Brian Y., Shamim Mollah, Dylan Skola, Michelle Dow, Chun-Nan Hsu, and Hannah Carter. 2018. “Creating a Scalable Deep Learning Based Named Entity Recognition Model for Biomedical Textual Data by Repurposing BioSample Free-Text Annotations.” bioRxiv. doi:10.1101/414136.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. ’t Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater,

George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3 (March). The Author(s): 160018.

Zhang, Chao, Kyle Cleveland, Felice Schnoll-Sussman, Bridget McClure, Michelle Bigg, Prashant Thakkar, Nikolaus Schultz, Manish A. Shah, and Doron Betel. 2015. “Identification of Low Abundance Microbiome in Clinical Samples Using Whole Genome Sequencing.” Genome Biology 16 (November): 265.

# CHAPTER 3: Creating a scalable deep learning based Named Entity Recognition Model for biomedical textual data by repurposing BioSample free-text annotations

## 3.1 Abstract

Extraction of biomedical knowledge from unstructured text poses a great challenge in the biomedical field. Named entity recognition (NER) promises to improve information extraction and retrieval. However, existing approaches require manual annotation of large training text corpora, which is laborious and time-consuming. To address this problem we adopted deep learning technique that repurposes the 43,900,000 Entity- free-text pairs available in metadata associated with the NCBI BioSample archive to train a scalable NER model. In our future development, this NER model can potentially assist in biospecimen metadata annotation by extracting named-entities from user-supplied free-text descriptions.

We evaluated our model against two validation sets derived both from the BioSamples, namely data sets consisting of short-phrases from specific entities (e.g. cell type, genotype, disease, etc) and long sentences from composite entities (e.g. TITLE and DESCRIPTION fields). We achieved an accuracy of 93.29% and 93.40% in the short-phrase validation set and long sentence respectively. This result suggests that we can utilize the short-phrases from BioSample to extract the more specific entities from long sentences made up of composite entities.

## 3.2 Introduction

Named Entity Recognition (NER) is an important task in the biomedical Natural Language Processing (NLP) domain. For example, NER can extract entities such as diseases, species, treatments, genes or geographical locations from biomedical free-text. A well constructed NER allows efficient document categorization and thus improve document retrieval accuracy.

Using NLP to automate the extraction of entity labels from biomedical textual data continues to be an area of active research, which often involves first curating a large terminology or annotating a large corpus for text model generation. For example, biomedical NLP engines like MetaMap (Aronson, 2001) captures the text variations using the Unified Medical Language System (UMLS), which integrated the vocabularies and their lexical variations from more than 60 biomedical vocabulary (Bodenreider, 2004). Recently, cTakes (Savova et al., 2010) and TaggerOne (Leaman and Lu, 2016) have been successfully applied in biomedical research settings for automated information extraction by using a mix of machine-learning, rule-based algorithms, and corpus building. However, the body of biomedical knowledge is constantly increasing in size as well as language complexity (Huang and Lu, 2016). This trend means that new biomedical text will be composed using words, sentence structures or entity types that were unseen by the model at the time of corpus construction. Thus, manual corpus annotation is slow to adapt to such changes as corpus generation is often limited by the high financial cost of hiring expert curators.

Another approach that has been adopted involves extracting a set of predefined terms from free-text to increase the relevancy of retrieved data (Galeota and Pelizzola, 2017; Barrett et al., 2013; Shah et al., 2009), where the terms and their relationships are often curated in the form of an ontology. However, the maintenance of coherency in an ontology is often difficult and also requires manual curation. In summary, existing approaches for NER do not scale well because of their dependence on manual curation.

In recent years, deep learning techniques have replaced much of the pre-processing steps required to extract knowledge from free-text. Some of the techniques include 1) word embedding to represent the concepts of words in dense format (Mikolov et al., 2013), and 2) recurrent neural network (RNN) which utilizes Long Short-Term Memory (LSTM) cells to capture the dependency of words (Sutskever et al., 2014). Recently, studies have shown that deep learning was capable of performing NER with high accuracy in biomedical texts (Zhu et al., 2018; Wu et al., 2017). However, their approaches require a large training corpus.

Here, we utilized the millions of BioSample (Barrett et al., 2012) metadata annotations hosted by NCBI to train a deep-learning based NER model. The BioSample metadata is natively encoded as Entity- free-text pairs and contains over 1,000,000 sample annotations spanning over 100,000 studies, making it a great resource for training a deep-learning-based NER model with minimal and "distant supervision" (Peng et al., 2016; Mintz et al., 2009). The BioSamples are comprised of the NCBI primary archives, including the Sequence Read Archive (SRA) (Kodama et al., 2012) and the database of Genotypes and Phenotypes (dbGAP) (Mailman et al., 2007). Given the community driven nature of BioSample reporting, we believe that the NER model can be kept up to date with the evolving biomedical vocabulary by simply retraining with the latest BioSample. Therefore, we explored the potential of repurposing the vast amount of BioSample annotations for training an NER model.

The first challenge we encountered in repurposing the BioSample was that the entity labels were not consistently annotated, yielding limited data coverage, which we solved by using the word embedding to merge the synonymous entity labels to increase the sample coverage in model training. We then adopted deep learning techniques to repurpose these BioSamples to train an NER model (**Figure 3.1A**), by first generating a short phrase classification model with an accuracy of 93.29%. Subsequently, we constructed an NER model for complete sentences from the vague

description fields derived from BioSample (e.g. TITLE and DESCRIPTION fields), by utilizing the scores emitted by the short phrase classification and N-gram model (Brown et al., 1992). We achieved an accuracy of  93.4% for this NER model. We also compared this NER model with MetaMap (Aronson, 2001).

## 3.3 Method and materials

***BioSample data landscape***. Each BioSample record is associated with a single biospecimen and has the metadata encoded in the form of entity- free-text pairs (Note that we alias the BioSample attributes as entities here). The BioSample annotation records were downloaded from the NCBI FTP website (ftp.ncbi.nlm.nih.gov/sra/reports/Metadata/) on May 15, 2018. The XML-encoded sample files were parsed into a python pickle object to allow simple data loading. Only ASCII characters from the BioSample are retained and python package spaCy was used for word tokenization with default parameters. We retrieved 2,921,722 BioSample records (SRA BioSample Symbol: SRS), spanning over 106,110 studies. The resulting python pickle object for SRS was 1.4GB. An example record is shown in **Figure 3.1A**. Details for BioSample annotation structure can be found at https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/ and https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/.

The BioSample entities are derived from both the submitters and from the official BioSample website. The free-text and the submitter driven nature of the BioSample entities suggest that it is diverse (n=19,996) as they recorded the experimental conditions from over 100,000 studies. The retrieved biospecimen records consist of 43,907,007 Entity- free-text pairs. The number of unique text entries associated with each entity scales with the number of SRS records, suggesting that the BioSample free-text annotations are very diverse for each entity. The 145 BioSample entities are associated with over 10,000 biospecimen records and the top 30 most frequently used entities are listed in (**Figure 3.1B**), providing a large volume of annotation data available for training a deep learning based phrase classification model. Moreover, composite entities made up of long sentences like DESCRIPTION and TITLE are among the top 10 most frequently used entities in BioSample (**Figure 3.1B**). These fields have a mean length of 46.10 characters, 3.19-fold longer (**Figure 3.2**) than the rest of the free-text fields (mean=14.46 characters, 95% confidence interval:

1.0 - 65.0 characters, **Figure S3.2**). These composite entities made up of specific experimental entities like source name, age, sex, etc. This begs the question of whether we can reclassify those annotations using an NER trained by the more atomic entities.

***Vectorization of biomedical words to expand data coverage in NER model training***

***Word embeddings retrieval***. Traditionally, words have been represented by a sparse hot-one encoding scheme in NLP. However, this trend changed when word embedding was introduced (Mikolov et al., 2013). In principle, the semantic similarity of words should be representable by geometric distances between trained word embeddings. This approach has the advantage of not requiring the laborious process of pre-defining the meaning of words by hard-coding semantics for each word or generating rules for part-of-speech tagging. For this project, we used a published biomedical word embedding model (Chiu et al., 2016) that was trained on the entire PubMed, PMC and Wikipedia text corpora, with a total of 5,443,656 word vectors where each word is represented by 200 features.

***Evaluation of semantic similarities at word, sentence, and entity resolution using word embedding***. The first validation of the utility of word embedding for this task was in the automatic identification of semantically similar words. The high number of word vectors enabled capture of word variations. For example, the word embedding model grouped related sequencing keywords in the PCA (Principal Component Analysis) space (**Figure 3.3**; variance explained: PC0: 34.2%; PC1: 20.4% ).

The second utility of word embedding we found was in combining semantically similar entities to increase BioSample coverage for model building and validation, as it has been recognized that there is a lack of standardization in metadata annotation (Brazma et al., 2001). To summarize and represent a free-text sentence, a sentence embedding vector can be generated by taking the feature-wise mean of the embedding vectors of the words with the free-text (Iyyer et al.,

2015). For example, sex is synonymous with gender and tissue is synonymous with cell type. When we randomly drew 100 biospecimen records with sex, gender, tissue and cell type BioSample entities, the sentence embedding vectors were able to group the free-texts according to their semantics (**Figure 3.4**). To summarize at the entity level, we further reduced the sentence level embeddings to entity level embeddings using the same method. The entity embedding vectors of the 30 most commonly used BioSample entities cluster by cosine similarity, which further shows higher embedding similarities for the more semantically similar entities (**Figure S3.3**).

***Expansion of NER training data coverage using entity embedding similarities***. The entities selected for our NER model are the six commonly used metadata attributes: Species, Genotype, Disease state, Cell type/tissue, Geographical Location, and Treatment/Conditions. The selection considers the semantic similarity and data size of the attributes available in BioSample. We first incorporated the corresponding official BioSample attributes: SCIENTIFIC_NAME, genotype, disease, cell type, geo_loc_name, treatment. Then, to increase the sample coverage for each entity class in NER training, we identified all the entities with a high cosine similarity (>0.8) with the corresponding official BioSample entity. We observed that all those retrieved entities deemed synonymous from the label names except "cell description" which became grouped with "disease state" (**Table S3.1**), suggesting that most of the entities in the BioSample with high embedding similarities can be potentially merged to reduce the complexity of the database and increase sample size for each equivalent class. For example, the sample coverage of genotype entity increased by 62.65% after merging the BioSample entities with high embedding similarities.

***NER model training and validation***. An NER model takes in free-text of any length and extracts the biomedical entities in phrases. Traditional entity recognition models require the corpus to be in complete sentences with entity labels annotated in order for the model to segment and identify phrase boundaries, as exemplified by the corpus format of the CoNLL shared task (Tjong

Kim Sang, 2002), one of the most widely-used data set for NER research. However, this approach is not useful for the BioSample metadata which is encoded as Entity- free-text pairs. Therefore, we hypothesized that we could first train a short phrase entity classification model and then use an n-gram approach to extract named entity segments in longer sentences by identifying n-grams with high emission probability (**Figure 3.5**).

***Short phrase entity classification model construction.*** In order to train the short phrase entity classification model, we retained only text comprising 2 to 7 words. The upper bound of 7 words corresponds to the 95th percentile of the distribution of phrase lengths in a highly curated general purpose medical vocabulary called the NCI Thesaurus (Sioutos et al., 2007/2). This filtering also has the advantage of limiting the number of parameters in the Recurrent Neural Network (RNN) model. To build and validate the deep learning based NLP model for short phrase entity classification, we first randomly split the training and validation cohorts in a 4:1 ratio by studies, to maximize training-validation set independency. The deep NLP model is trained based on a bidirectional RNN architecture with LSTM cells (biLSTM) (Graves and Schmidhuber, 2005).

For both the training and test data, we first converted the biospecimen free-text to a sentence by word ID matrix, where each row in the matrix consists of a sequence of word embedding IDs. Then, the model was constructed and trained with the following layers: An embedding layer to convert word embedding IDs to word vectors.A bidirectional layer with a total of 64 hidden units, and a dropout rate of 0.5. A dense, fully connected layer with logistic activation function for outputting the probability score used for classification. The number of neurons in this layer is the same as the number of entities. The Adam optimizer (Kingma and Ba, 2014) with categorical cross-entropy as loss function was used to compile the deep learning model. This biLSTM model was constructed and trained using keras v2.7.1 with tensorflow v1.8.0 backend on

a single machine with 48 physical cores with four Intel(R) Xeon(R) CPU E5-4650 v3 @ 2.10GHz, with a learning rate of 0.001 and batch size of 100. The training completed within 1 hour.

***Long sentence NER model construction and validation.*** Each free-text input is segmented into sentences using commonly used separators like ':' , '.' , ',' . In total, 179 stop words from the python package NLTK were removed from the input sentence. For each sentence, the non-alphanumeric characters were removed. To generate entity scores for each n-gram, the algorithm scans for entities among all phrases that match phrase lengths specified during model construction, by applying the short phrase classification model on each n-gram with n ranges from 2 to 7 in the sentence. Only n-grams with at least 2 tokens matching with word embedding IDs are retained. The tokens were then converted to word embedding IDs and became input to the deep NER model for short phrase entity classification. Therefore, each n-gram will be associated with a vector of entity scores emitted by the logistic activation function from the last layer of the biLSTM, ranging from 0.0 to 1.0.

We also estimated the baseline neural net emission entity scores by executing the prediction function without any input. All the n-grams from emission vectors with absolute sum difference < 0.01 with baseline emission are zeroed. For overlapping n-grams that have the same entity, only the n-gram with the highest probability score will be retained.

***Manual curation set.*** To compare our NER model against manual curation, we manually curated 185 sentences selected from 100 randomly drawn BioSample records with a DESCRIPTION entity using Dataturk (https://dataturks.com). For each sentence, we highlighted the token segments that described the entities selected in model construction. We then downloaded the annotated data in JSON format and compared it against the prediction results. We also generated a second validation cohort by manually curated 327 sentences selected from 261 randomly drawn BioSamples to further evaluate the model performance. The validation sample

size in our cohort is similar to the validation sample size used by Bernstein et al. when they evaluated their normalization of the SRA metadata (Bernstein et al., 2017).

***MetaMap set***. We compared the performance of our NER model against MetaMap from the National Library of Medicine, which has been adopted by GIANT (Greene et al., 2015) for automating biospecimen annotation extraction. We used the online MetaMap service (URL: https://ii.nlm.nih.gov/Batch/UTS_Required/MetaMap.shtml) to extract the terms from the validation cohort. In the case where the same text segment is mapped onto multiple terms, only the term with the highest MetaMap score is retained. In order to match the UMLS semantic types generated from MetaMap to our entities, we mapped the UMLS semantic types to UMLS groups using the table from the MetaMap website: https://MetaMap.nlm.nih.gov/Docs/SemGroups_2013.txt.

## 3.4 Results

***Short phrase entity classification model performance***. The short phrase entity recognition model was able to classify 93.29% percent of the data correctly in the validation cohort (See Methods). All the entities were able to achieve an ROC-AUC (**Figure 3.6A**) and an accuracy (**Figure 3.6B**) over 90% in the validation cohorts, suggesting a good sensitivity and specificity. Next, we also observed a high F1 score of 92.14% across all the entities, suggesting our the model also has high precision and recall over all the classes. Also, the accuracy of the short phrase entity recognition model improved when trained with more data (**Figure 3.7**), confirming the utility of using the ever-growing source of community-contributed BioSample data to train a more accurate NER model.

***NER performance in long sentence***. Since BioSample metadata is in the format of entity-free-text pairs as opposed to a fully annotated corpus, we evaluated the possibility of using an n-gram approach to segment the text. For each n-gram, we assign the entity with the highest short phrase classification score. We compared the performance of the model against manual curation and the existing tool MetaMap (Aronson, 2001). We manually curated 100 free-text entries from samples with the DESCRIPTION field, with a total of 185 sentences where 139 sentences had at least one entity extracted in curation. **Figure 3.8** shows an representative example of the NER annotation, which shows both the capability and the limitation of our method, our method could recognize “liver cells” as a cell type and “C57 BL6” as the mouse genotype like a regular NER without prior expert curations. None of the less, we can also foresee multiple limitations from this example. The first limitation we foresee is the span annotation inaccuracy. The entity “treatment” which is often open-ended regarding the most fitting spans, our NER only pick up certain recurrent keywords without annotating a complete phrase. Therefore, we evaluated both the entity membership recovery and the exact span recovery capability of our NER.

The model yielded an accuracy of 93.4% in entity memberships recovery for each sentence. We then compared the performance of the model against MetaMap. For the task of biological entity recognition, our NER model had superior performance in terms of precision, sensitivity and specificity and also an F1-score (**Table 3.1**). To further evaluate our model, we have also a generated a second independent validation cohort with a total of 327 sentences with at least one entity extracted in curation. The second validation cohort also yielded a high F1 score of 0.848, similar to the F1 score of 0.876 in our initial validation cohort, confirming that our NER model is robust. Nonetheless, the predicted spans of each entity from our model has discrepancy with the curated spans in the validation data, where only 62.9% of the curated intervals overlap with predicted data (**Figure S3.4**).

The lower precision and recall in MetaMap is likely due to the reliance on a set of hand-crafted rules to capture the text variations (Aronson and Lang, 2010). Here, we use a submitter-based resource which enable the model to be accurate and kept up-to-date without manual curation. For example, our model was able to accurately extract phrases like “germline sp1” as Genotype, while MetaMap mapped “germline” as a cell type “Germ Line” and did not recognize “sp1” as Genotype related. Also, instead of recognizing “MRG15 null” (DRS033379) as a single entity of Genotype like our model, MetaMap recognized the word “MRG15” as a Gene and “null” as an Qualitative concept. Moreover, words that are missing in the supplying terminology cannot be mapped with MetaMap. For example, MetaMap fails to capture the new word “sc-197” (ERS396144), which is an antibody that was first developed in 2010 and is not recorded in the latest UMLS.

## 3.5 Discussion

Here, we trained a deep learning based NER model by repurposing the vast amount of BioSample metadata in NCBI, available as entity- free-text pairs. We first showed that the word embeddings were able to group the free-text annotations at various resolutions, i.e. word, sentence and entity level. This allowed us to utilize entity embeddings to merge synonymous entity labels to increase sample coverage for short phrase entity classifier training. The short phrase entity classification model was effectively able to extract entities from phrases in the validation set. We then used this model to extract named entities from long sentences. We were able to achieve a higher accuracy using our NER model when compared with existing method MetaMap (Aronson, 2001).

Existing methods require a large annotated data corpus that is difficult to come by. For example, existing biomedical NER has relied on laborious corpus curation and annotation of the entities in each sentence. Expert curation is costly in terms of both time and monetary expenses, and may suffer from curation bias. In contrast, our approach can incorporate new vocabulary and entity definitions without the need for human curation. The word embeddings (Chiu et al., 2016) used here were generated by PubMed articles and Wikipedia, and entity-free-text pairs were generated by the biomedical research community. We then utilized neuron emission strengths from the trained NER for phrase segmentation. This entire process was free of any manual expert curation, in large thanks to the large volume of BioSamples generated by the research community. As a result, this model can be readily extended in vocabulary and NER capability by simply incorporating more training examples or new embeddings as they become available.

Furthermore, our model has a concise code base. For example, the code base in cTakes consists of 1,404 java files with a total of 234,388 lines of code, and the code base in MetaMap consist of 218 files with 260,586 lines of code, while our model can be fully specified in less than

200 lines of code. Our approach eliminates the need for stemming, part of speech tagging (POS) and dependency parsing. Thus, deep learning simplifies model construction and provides good performance for automated biological entity recognition.

We acknowledged that there are limitations to our approach. For example, there is no professional curator to ensure the correctness of the BioSample entries which may have contributed to the lower accuracy of our NER prediction. In addition, we incorporated a limited number of entities into the model. Further evaluation is merited to determine whether adopting a Convolution Neural Network (CNN), or adding Conditional Random Fields (Zheng et al., 2015) or attention layers (Luong et al., 2015) will further improve the model. Also, we did not include any numerical entities due to the lack of availability of annotations (only the age entity had more than 10,000 biospecimen annotations). Also, merging the synonymous entities using the embedding method could introduce noise in the model if the merged entities are not equivalent. Lastly, though we derived two cohorts of validation data, the total sample size of our evaluation cohort remains relatively small, which requires curating a larging validation cohort.

In the future, we are also interested in better understanding the confounding relationships between the entities in the BioSample archive and developing algorithms to identify synonyms entities, possibly by using community finding algorithms like Clixo (Kramer et al., 2014). We are also interested in evaluating how increasing the sample size using the embedding methods can affect the accuracy and the generalizability of the model. The other aspect we are excited to pursue would be creating a web-based metadata reformatting platform with our NER as the backend, where this tool can suggest the entities that the users should annotate based on the text from the manuscript.

## 3.6 Figures

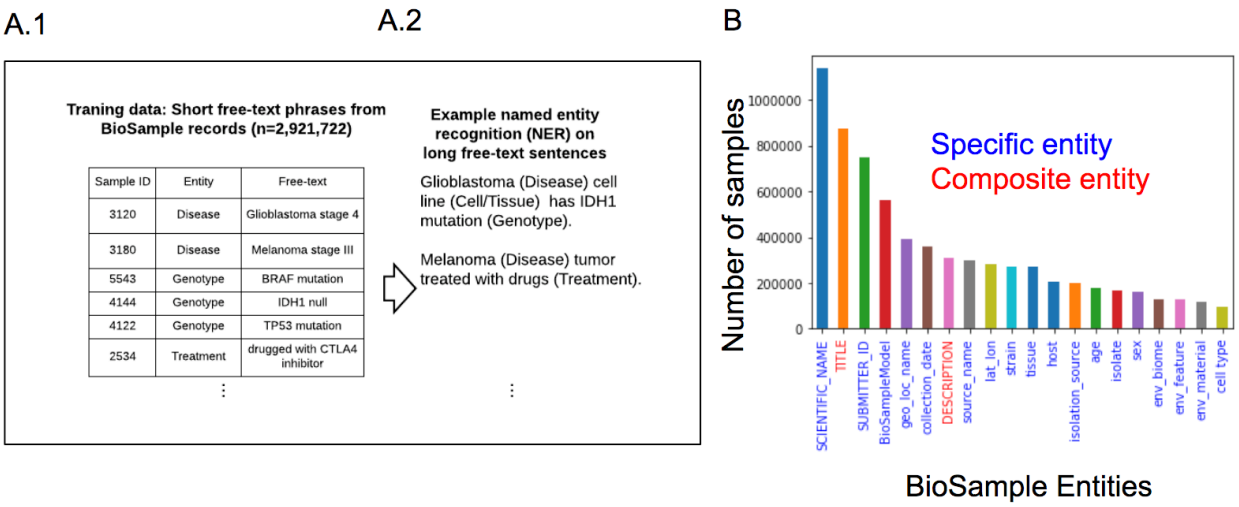



**Figure 3.1. Repurposing public biospecimen annotation data for NER training.**
(**A**) Depiction of training NER model using pre-annotated Entity- free-text pairs available from public biospecimen annotation data (BioSamples) from NCBI (**A.1**) Example of Entity- free-text pairs from BioSamples. In this example, the free-text phrase "Glioblastoma stage 4 system" is a "Disease" entity. (**A.2**) Expected results of an NER model recognizing biomedical concepts from sentences. (**B**) Histogram of the 30 most frequently used entities (x-axis) available in the current set of BioSamples. These atomic named entities (blue labels) can be used to extract concepts from composite entities TITLE and DESCRIPTION (red labels).

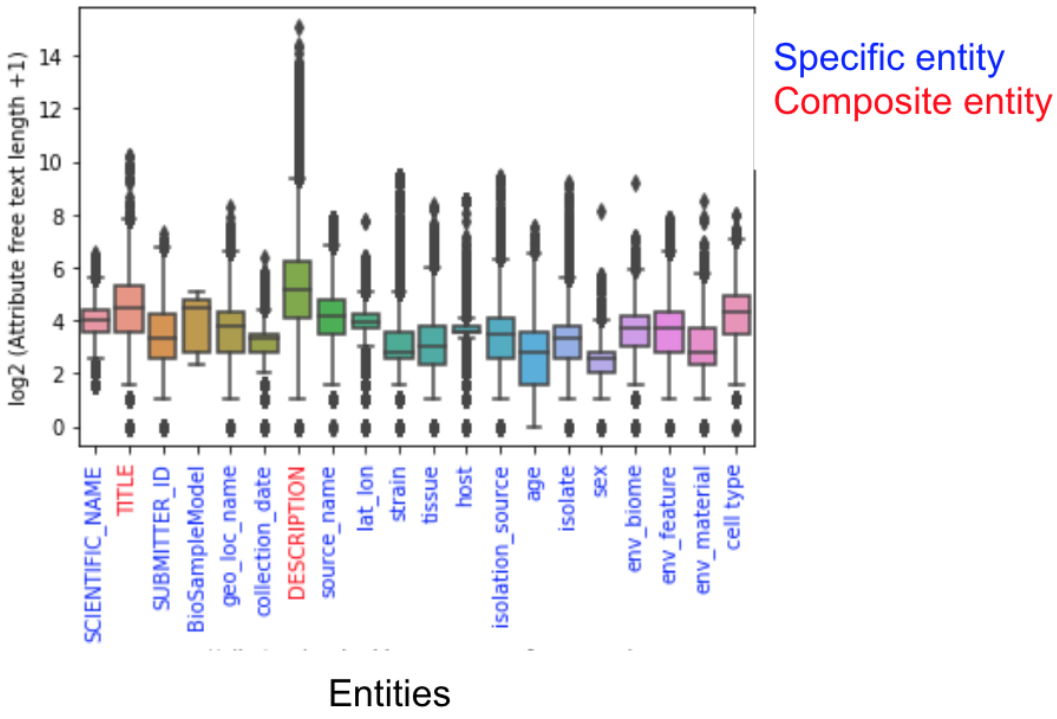



**Figure 3.2. Text length distribution of top entities.**

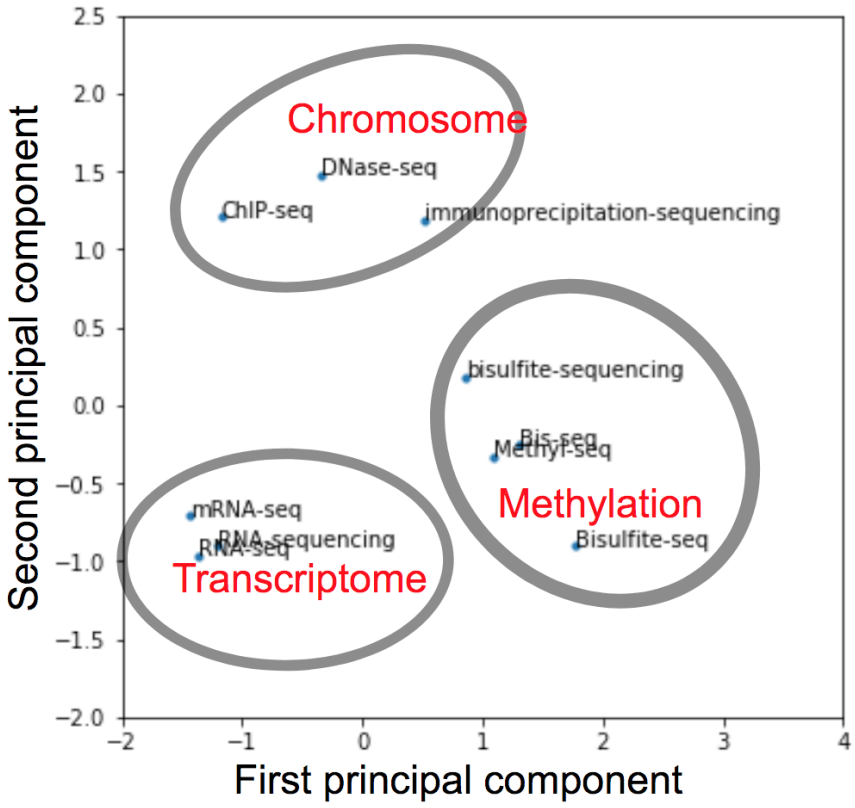



**Figure 3.3. Word embeddings group related terms in PCA space**

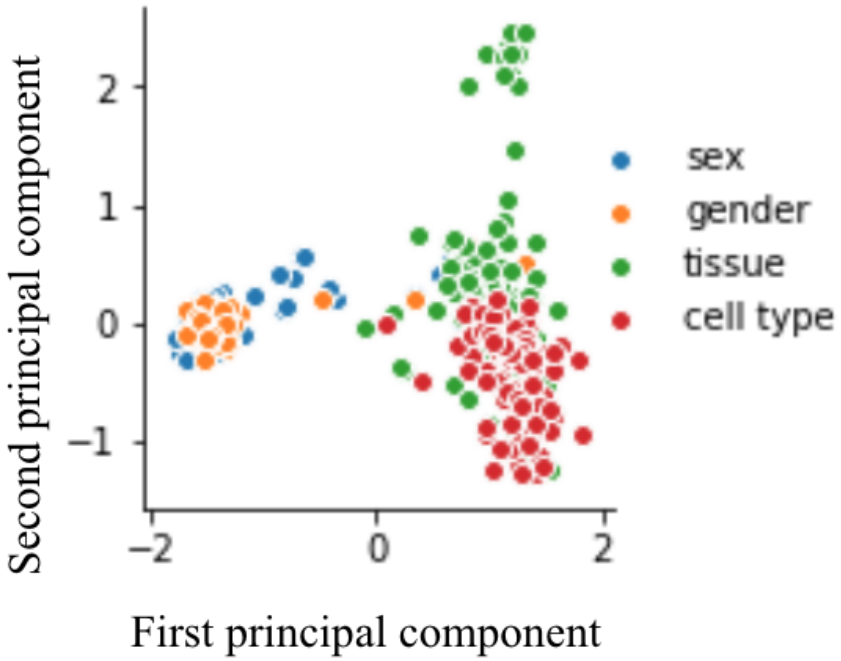



**Figure 3.4. Native sentence embeddings recover reasonable grouping.**

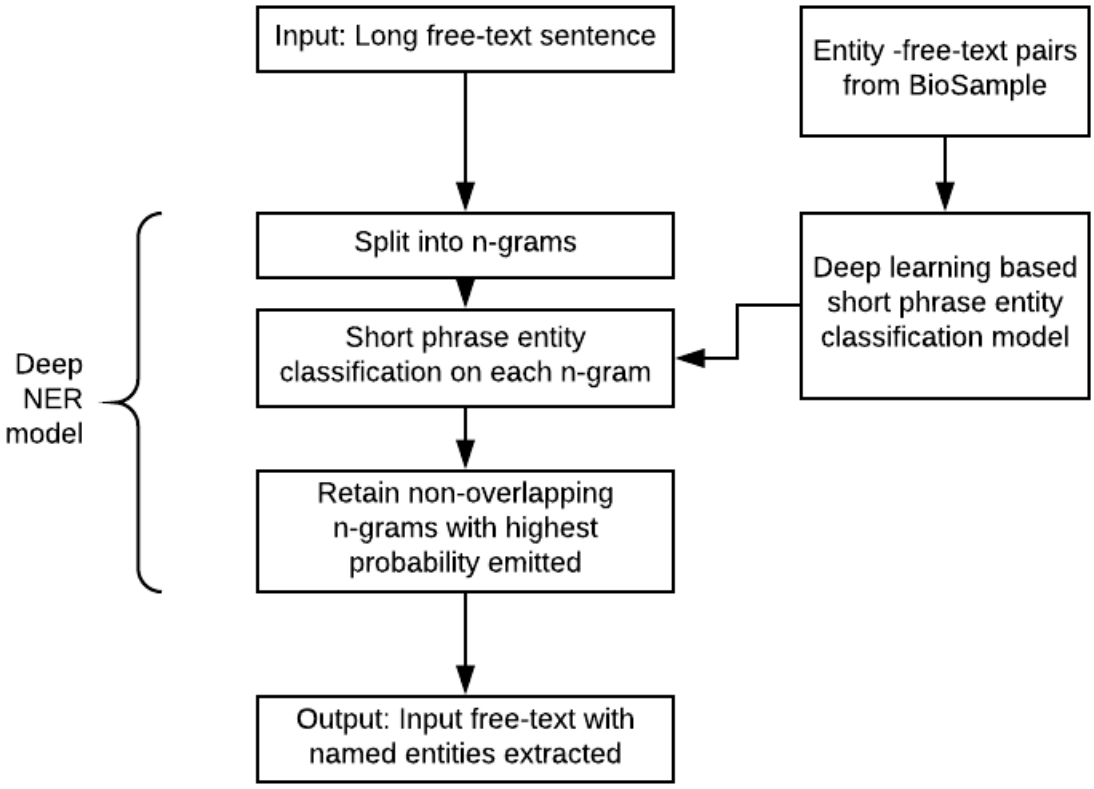



**Figure 3.5. Depiction of our NER model**

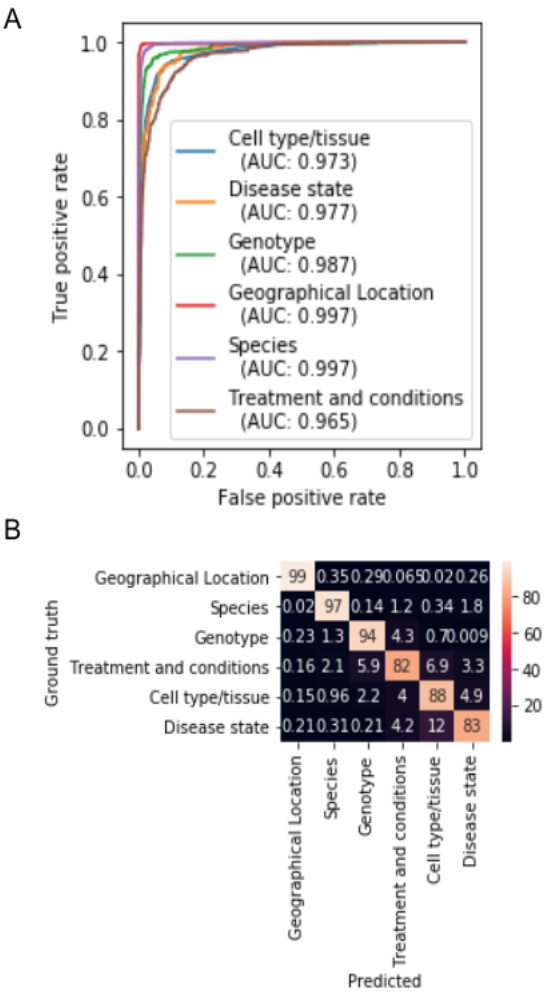



**Figure 3.6. Short phrase classification model performance.**
(A) ROC-AUC of entities (B) Contingency table quantifying percentage of entities predicted correctly in validation cohort.

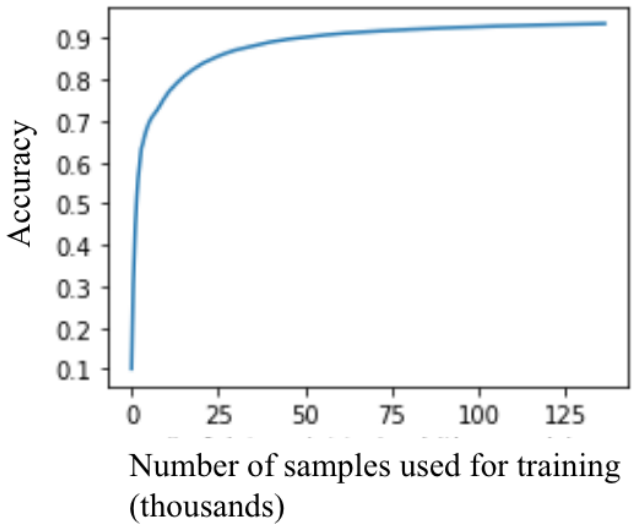



**Figure 3.7. Performance scale with the volume of training  data.**

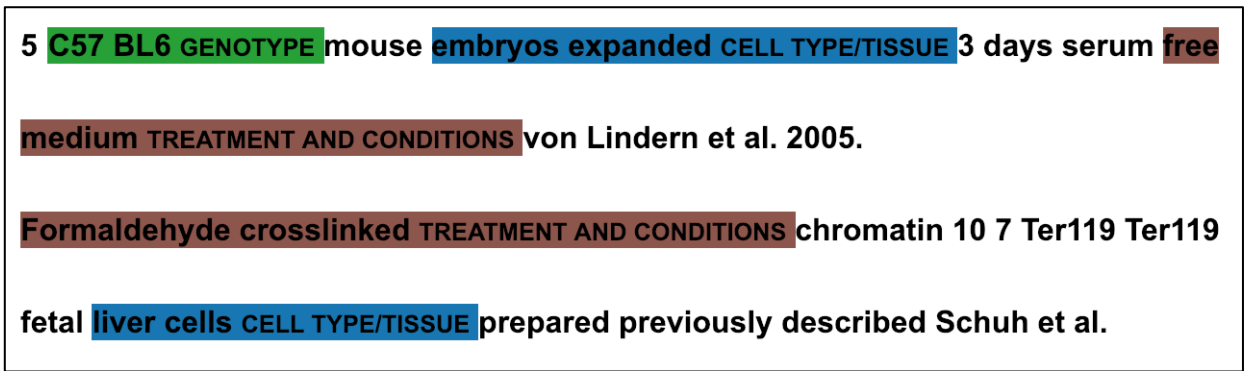



**Figure 3.8. Example description annotation**
This example description annotation was collected from BioSample with identifier ERS215384. Each region is color coded based on the entity type classified by the NER model.

## 3.7 Tables

**Table 3.1. NER performance comparison with MetaMap.**

| | Deep bio NER | Metamap |
|---|---|---|
| **Sensitivity /recall** | 0.932 | 0.654 |
| **Specificity** | 0.941 | 0.850 |
| **Precision** | 0.827 | 0.739 |
| **F1-score** | 0.876 | 0.694 |

## 3.8 Supplemental Figures

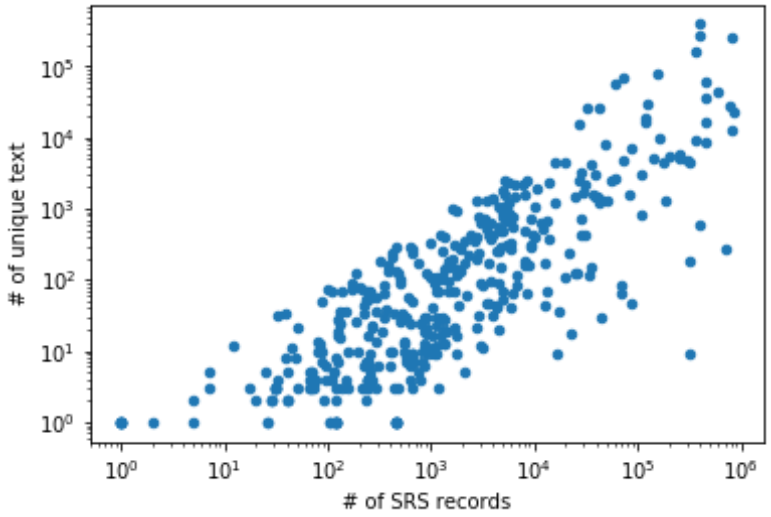



**Figure S3.1. The number of unique free text annotations scales with the number of BioSample records for each BioSample entity.**

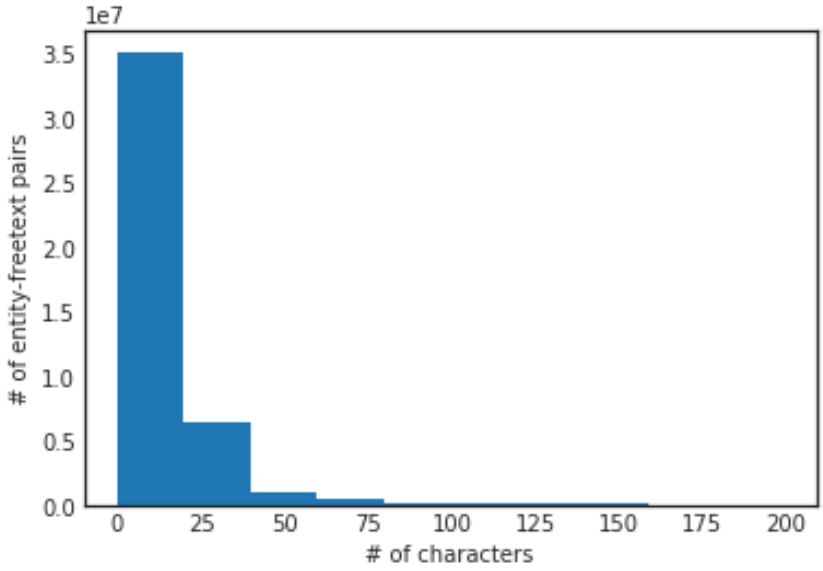



**Figure S3.2. Distribution of free text length in biospecimen annotations**
Distribution of free text length (x-axis) over SRA entity- free-text pairs (y-axis)

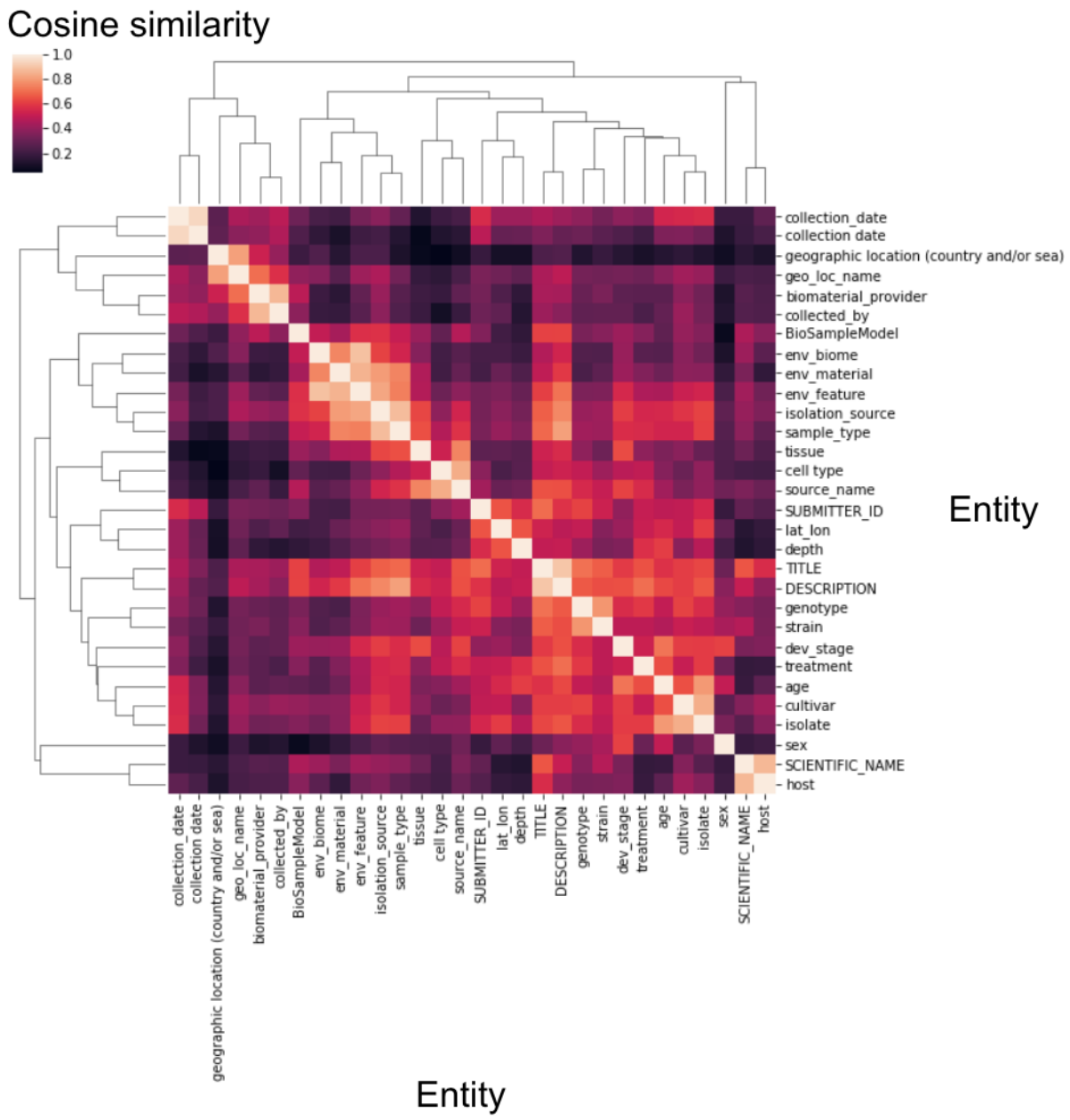



**Figure S3.3. Entity embedding recovers biospecimen entity similarities.**

## 3.9 Supplementary Table

**Table S3.1. Characteristics and mapping statistics of the SRA sequencing runs.**

| Entity class name | BioSample entity name | Cosine similarity | # of samples per BioSample entity | # of samples (aggregated) | Fold increase |
|---|---|---|---|---|---|
| Species | SCIENTIFIC_NAME | 1.00 | 1,136,856 | 1,539,081 | 1.35 |
| | organism | 0.98 | 29,037 | | |
| | Organism | 0.90 | 2,894 | | |
| | host scientific name | 0.86 | 9,909 | | |
| | Species | 0.85 | 578 | | |
| | host | 0.84 | 205,511 | | |
| | specific host | 0.83 | 11,114 | | |
| | host_scientific_name | 0.83 | 4,297 | | |
| | host organism | 0.83 | 372 | | |
| | nat-host | 0.82 | 1,516 | | |
| | specific_host | 0.81 | 14,088 | | |
| Genotype | genotype | 1.00 | 75,566 | 122,909 | **1.63** |
| | genotype/variation | 0.97 | 28,730 | | |
| | plant genotype | 0.88 | 195 | | |
| | mutant | 0.85 | 314 | | |
| | mutation | 0.83 | 428 | | |
| | phenotype | 0.82 | 13,892 | | |
| | host_genotype | 0.80 | 3,784 | | |
| Disease state | disease | 1.00 | 21,654 | 34,239 | **1.58** |
| | tumor type | 0.87 | 691 | | |
| | diagnosis | 0.86 | 3,351 | | |
| | disease state | 0.85 | 3,715 | | |
| | DiseaseState | 0.83 | 173 | | |
| | cancer type | 0.82 | 311 | | |
| | tumor | 0.82 | 355 | | |
| | clinical history | 0.82 | 280 | | |
| | disease status | 0.82 | 2,077 | | |

| | | | | | |
|---|---|---|---|---|---|
| | cell description | 0.80 | 1,632 | | |
| Cell type/tissue | cell type | 1.00 | 94,819 | 417,924 | 4.41 |
| | cell description | 0.94 | 1,632 | | |
| | cell_type | 0.93 | 18,162 | | |
| | source cell type | 0.93 | 106 | | |
| | cell types | 0.91 | 160 | | |
| | source_name | 0.91 | 297,941 | | |
| | cell-type | 0.89 | 286 | | |
| | CellType | 0.88 | 342 | | |
| | cell subtype | 0.88 | 1,220 | | |
| | biomaterial_type | 0.86 | 182 | | |
| | progenitor cell type | 0.86 | 169 | | |
| | tissue/cell type | 0.85 | 783 | | |
| | DIFFERENTIATION_STAGE | 0.84 | 170 | | |
| | cell | 0.83 | 1,616 | | |
| | differentiation status | 0.82 | 111 | | |
| | cell line source | 0.81 | 225 | | |
| Geographical Location | geo_loc_name | 1.00 | 389,901 | 509,997 | 1.31 |
| | geographic location | 0.96 | 13,640 | | |
| | geo loc name | 0.94 | 704 | | |
| | geographic location (country and/or sea, region) | 0.92 | 7,920 | | |
| | geographic location (country and/or sea,region) | 0.87 | 12,833 | | |
| | country | 0.86 | 24,413 | | |
| | geographic location (country and/or sea,region) | 0.84 | 539 | | |
| | birth_location | 0.83 | 3,182 | | |
| | geographic location (country and/or sea) | 0.81 | 56,576 | | |
| | Geo_loc_name | 0.80 | 289 | | |
| Treatment and conditions | treatment | 1.00 | 73,041 | 90,516 | 1.24 |
| | treated with | 0.92 | 3,026 | | |
| | treatment protocol | 0.91 | 583 | | |
| | drug treatment | 0.89 | 977 | | |

| | | | | | |
|---|---|---|---|---|---|
| | **agent** | **0.89** | **2,570** | | |
| | **stimulated with** | **0.86** | **247** | | |
| | **protocol** | **0.85** | **2,769** | | |
| | **Treatment** | **0.84** | **3,591** | | |
| | **sample group** | **0.82** | **1,176** | | |
| | **cell treatment** | **0.82** | **231** | | |
| | **experimental condition** | **0.82** | **187** | | |
| | **culture conditions** | **0.81** | **391** | | |
| | **culture condition** | **0.80** | **1,220** | | |
| | **treatment group** | **0.80** | **405** | | |
| | **sample treatment** | **0.80** | **102** | | |

## 3.10 Acknowledgement

We thank all members of the Carter and Ideker lab for scientific feedback and comments. This work was funded by NIH grants DP5-OD017937, RO1 CA220009 and a CIFAR fellowship to H.C. We do not have any conflict of interest with our current funding source.

Chapter 3, in part, is being prepared for submission and in full, is a reformatted reprint of the material as it appears as "Creating a scalable deep learning based Named Entity Recognition Model for biomedical textual data by repurposing BioSmaple free-text annotations" in *bioRXiv*, 2018 by Brian Tsui, Shamim Mollah, Dylan Skola, Michelle Dow, Chun-Nan Hsu, Hannah Carter. The dissertation author was the primary investigator and author of this paper.

## 3.11 References


Aronson, Alan R., and François-Michel Lang. 2010. “An Overview of MetaMap: Historical Perspective and Recent Advances.” Journal of the American Medical Informatics Association: JAMIA 17 (3): 229–36.

Aronson, A. R. 2001. “Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program.” Proceedings / AMIA ... Annual Symposium. AMIA Symposium, 17–21.

Barrett, Tanya, Karen Clark, Robert Gevorgyan, Vyacheslav Gorelenkov, Eugene Gribov, Ilene Karsch-Mizrachi, Michael Kimelman, Kim D. Pruitt, Sergei Resenchuk, Tatiana Tatusova, Eugene Yaschenko, and James Ostell. 2012. “BioProject and BioSample Databases at NCBI: Facilitating Capture and Organization of Metadata.” Nucleic Acids Research 40 (Database issue): D57–63.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. 2013. “NCBI GEO: Archive for Functional Genomics Data Sets—update.” Nucleic Acids Research 41 (D1). Oxford University Press: D991–95.

Bernstein, Matthew N., Anhai Doan, and Colin N. Dewey. 2017. “MetaSRA: Normalized Human Sample-Specific Metadata for the Sequence Read Archive.” Bioinformatics 33 (18): 2914–23.

Bodenreider, Olivier. 2004. “The Unified Medical Language System (UMLS): Integrating Biomedical Terminology.” Nucleic Acids Research 32 (Database issue): D267–70.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. 2001. “Minimum Information about a Microarray Experiment (MIAME)-toward Standards for Microarray Data.” Nature Genetics 29 (4): 365–71.

Brown, Peter F., Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. “Class-Based N-Gram Models of Natural Language.” Computational Linguistics 18 (4). Cambridge, MA, USA: MIT Press: 467–79.

Chiu, Billy, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. “How to Train Good Word Embeddings for Biomedical NLP.” In Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 166–74.

Galeota, Eugenia, and Mattia Pelizzola. 2017. “Ontology-Based Annotations and Semantic Relations in Large-Scale (epi)genomics Data.” Briefings in Bioinformatics 18 (3): 403–12.

Graves, Alex, and Jürgen Schmidhuber. 2005. “Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures.” Neural Networks: The Official Journal of the International Neural Network Society 18 (5-6): 602–10.

Greene, Casey S., Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, Boris M. Hartmann, Elena Zaslavsky, Stuart C. Sealfon, Daniel I. Chasman, Garret A. FitzGerald, Kara Dolinski, Tilo Grosser, and Olga G. Troyanskaya. 2015. “Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.” Nature Genetics 47 (6): 569–76.

Huang, Chung-Chi, and Zhiyong Lu. 2016. “Community Challenges in Biomedical Text Mining over 10 Years: Success, Failure and the Future.” Briefings in Bioinformatics 17 (1). Oxford University Press: 132–44.

Iyyer, Mohit, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. “Deep Unordered Composition Rivals Syntactic Methods for Text Classification.” In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1:1681–91.

Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” arXiv [cs.LG]. arXiv. http://arxiv.org/abs/1412.6980.

Kodama, Yuichi, Martin Shumway, Rasko Leinonen, and International Nucleotide Sequence Database Collaboration. 2012. “The Sequence Read Archive: Explosive Growth of Sequencing Data.” Nucleic Acids Research 40 (Database issue): D54–56.

Kramer, Michael, Janusz Dutkowski, Michael Yu, Vineet Bafna, and Trey Ideker. 2014. “Inferring Gene Ontologies from Pairwise Similarity Data.” Bioinformatics 30 (12): i34–42.

Leaman, Robert, and Zhiyong Lu. 2016. “TaggerOne: Joint Named Entity Recognition and Normalization with Semi-Markov Models.” Bioinformatics 32 (18): 2839–46.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. 2015. “Effective Approaches to Attention-Based Neural Machine Translation.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1508.04025.

Mailman, Matthew D., Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, Natalia Popova, Stephanie Pretel, Lora Ziyabari, Moira Lee, Yu Shao, Zhen Y. Wang, Karl Sirotkin, Minghong Ward, Michael Kholodov, Kerry Zbicz, Jeffrey Beck, Michael Kimelman, Sergey Shevelev, Don Preuss, Eugene Yaschenko, Alan Graeff, James Ostell, and Stephen T. Sherry. 2007. “The NCBI dbGaP Database of Genotypes and Phenotypes.” Nature Genetics 39 (10): 1181–86.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” In Advances in Neural

Information Processing Systems 26, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, 3111–19. Curran Associates, Inc.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky. 2009. "Distant Supervision for Relation Extraction without Labeled Data." Of the Joint Conference of the 47th …. dl.acm.org. https://dl.acm.org/citation.cfm?id=1690287.

Peng, Yifan, Chih-Hsuan Wei, and Zhiyong Lu. 2016. "Improving Chemical Disease Relation Extraction with Rich Features and Weakly Labeled Data." Journal of Cheminformatics 8 (October): 53.

Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. "Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications." Journal of the American Medical Informatics Association: JAMIA 17 (5): 507–13.

Shah, Nigam H., Clement Jonquet, Annie P. Chiang, Atul J. Butte, Rong Chen, and Mark A. Musen. 2009. "Ontology-Driven Indexing of Public Datasets for Translational Bioinformatics." BMC Bioinformatics 10 Suppl 2 (February): S1.

Sioutos, Nicholas, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. 2007/2. "NCI Thesaurus: A Semantic Model Integrating Cancer-Related Clinical and Molecular Information." Journal of Biomedical Informatics 40 (1): 30–43.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. "Sequence to Sequence Learning with Neural Networks." In Advances in Neural Information Processing Systems 27, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, 3104–12. Curran Associates, Inc.

Tjong Kim Sang, Erik F. 2002. "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition." In Proceedings of the 6th Conference on Natural Language Learning - Volume 20, 1–4. COLING-02. Stroudsburg, PA, USA: Association for Computational Linguistics.

Wu, Yonghui, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. 2017. "Clinical Named Entity Recognition Using Deep Learning Models." AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium 2017: 1812–19.

Zheng, Shuai, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. 2015. "Conditional Random Fields as Recurrent Neural Networks." In Proceedings of the IEEE International Conference on Computer Vision, 1529–37.

Zhu, Qile, Xiaolin Li, Ana Conesa, and Cécile Pereira. 2018. "GRAM-CNN: A Deep Learning Approach with Local Context for Named Entity Recognition in Biomedical Text." Bioinformatics 34 (9): 1547–54.

# EPILOGUE

This dissertation builds a foundation towards automating the processing of -omic data association with meta-data. The first chapter applied a simple count-based approach in variant identifications of over 250,000 human sequencing experiments. The second chapter then created a Jupyter-hub environment in which the users can go from finding, interpolating, accessing and reusing the read counts from over 900,000 public sequencing experiments in minutes. This method leverages the fact that count-based data are useful in sequencing data analysis and have the advantage of a smaller computational resource footprint. To improve the utility of our re-processed data to the bioinformatics community, I also extended the bioinformatics pipeline from the first chapter to incorporate more -omic layers. The third chapter then addressed the other remaining challenges of reusing the public -omic data, which is automatically reformating the -omic metadata. We repurposed the community-based biospecimen annotations to derive an NLP model that could understand the biomedical language without expert curation, which has the advantage of expanding its vocabulary over time with the updated BioSamples. Though this dissertation has many limitations and shortcomings as discussed in each chapter, it is interesting to ponder the possibility of a computational method to mimic the human bioinformatics research process by automatically correlating the re-processed big -omic data with the NLP extracted experimental conditions.