**Title:**  A deep learning NLP platform for microbiome metadata harmonization and dataset selection

**Contact Information:**

Co-PI:

Hannah Carter, PhD

Department of Medicine

9500 Gilman Dr. MC 0688

La Jolla, CA 92093-0688

[hkcarter@health.ucsd.edu](mailto:hkcarter@health.ucsd.edu)

Tel: (858) 822-4706

Co-PI:

Chunnan Hsu, PhD

Department of Neurosciences

9500 Gilman Dr. MC 0608

La Jolla, CA, 92039-0608

chunnan@health.ucsd.edu

Tel: (858) 822-3606

CMI Faculty Member:

Trey Ideker, PhD

Department of Medicine

9500 Gilman Dr. MC 0688

La Jolla, CA, 92093-0688

tideker@health.ucsd.edu

Tel: (858) 822-4558

**Executive Summary:** In large biological data repositories, such as Qiita, data reusability for metaanalysis combining multiple studies is impeded by differences in study annotations (e.g. study metadata). We have developed a natural language processing tool based on deep learning using word embeddings to automate metadata imputation and successfully applied it to improve harmonization of biosample annotations in the SRA. Here we propose to adapt this approach to impute and harmonize metadata across microbiome studies in Qiita. We will further develop a webtool to autofill metadata fields when new datasets are added. As further proof of benefit, we will evaluate the impact of metadata harmonization through a metaanalysis of IBD.

**Research Plan:** The Qiita database hosts thousands of studies describing microbiome sequences across diverse conditions and technologies, and continues to grow rapidly. This data resource provides an unprecedented opportunity to study microbiome dynamics across a wide variety of settings and diseases. Meta-analyses across multiple studies can improve power to detect biological signals when sample sizes of individual studies are resource limited. Identifying and combining datasets effectively requires accurate and consistent metadata describing the contents and contexts of independent studies. Although Qiita enforces extensive requirements for sequence metadata annotation including use of MIxS standards at the time of sample submission, some samples are added through crowd sourcing and do not always strictly adhere. Furthermore, annotations of human phenotypes are not covered under MIxS, and the corresponding metadata are much less consistent. This poses a barrier to performing effective metaanalysis of human disease phenotypes with Qiita. Here we propose to adapt our existing deep learning framework for metadata imputation to address these challenges. This proposal builds on the joint expertise of the Hsu lab, the Carter lab and Ideker lab. Our goals will be accomplished in the following three aims:

**Aim 1: Train a PredictMEE deep learning classifier for Qiita metadata harmonization**

We will leverage existing metadata associated with the many studies in Qiita to predict optimal attribute labels for existing entries. This will address metadata quality issues and support improved selection and combination of datasets for metaanalysis.

**Aim 2: Create a webtool frontend to support user metadata entry when adding datasets to Qiita**

We will develop a web tool that will autofill metadata based on a block of text, such as an abstract, to simultaneously streamline and normalize user metadata input at the time of dataset submission.

**Aim 3: Perform a metaanalysis of IBD using studies retrieved post-metadata harmonization**

We will perform a metaanalysis of IBD pre- and post-metadata harmonization to quantify the advantage of improved dataset selection and phenotype annotation alignment for various statistical analyses.

**Approach:** Advances in next-generation sequencing (NGS) technologies have led to the rapid accumulation of publicly available microbial sequencing datasets in a variety of online repositories. High quality metadata annotations for data hosted in these repositories are essential for research reproducibility, and for conducting fast, powerful and scalable meta-analyses in studying various microbiomes. However, the level of diversity in microbiome studies and a lack of standards for depositing data into these repositories has made metadata plurality and harmonization a grand challenge in microbiome research.

A primary goal of metadata standardization efforts in microbiome research has been to establish the minimum information about any (x) sequence (MIxS) (1) requirement for submission of sequencing data to a repository. However, this requirement is often not enforced by major repositories, leading to heterogeneous metadata, impeding integration of datasets for meta-analysis. Though Qiita (2) requires that the description of the work, relevant publications, collection and processing parameters for each sample, and relevant covariates be included upon submission of each dataset, significant legacy metadata remains poorly annotated.

We recently designed and implemented a Natural Language Processing (NLP) approach utilizing deep neural networks to analyze and improve the quality of metadata in the NCBI Sequence Read Archive (SRA) (3). We leveraged almost 44 million attribute-value pairs from SRA BioSample to train a scalable, recurrent neural network that predicts metadata attribute values via Named Entity Recognition (NER) (**Figure 1**). For SRA, the network was trained to classify short text phrases according to membership in each of 11 metadata categories and achieved an overall accuracy and area under the receiver operating characteristic (AUROC) curve of 85.2% and 0.977 respectively (**Figure 2**). We then applied our classifier to predict the values of the 11 metadata categories from the longer TITLE attribute of samples, evaluating performance on a set of samples withheld from model training. Prediction accuracies were high when extracting sample Species (94.85%), Condition/Disease (95.65%) and Strain (82.03%) from TITLEs, with lower accuracies and lack of predictions for other categories highlighting multiple issues with the current metadata annotations in BioSample.

Here we propose to adapt this NLP deep learning framework to address two major challenges facing the Qiita metagenomic dataset repository: Quality issues with 1) standard metadata fields arising from differences in how users upload data to the database and 2) non-standard metadata attributes related to human phenotypes.

Aim 1: Train a PredictMEE deep learning classifier for Qiita metadata harmonization

This aim will adapt our existing NLP deep learning tool to address metadata quality issues specific to Qiita. We have already downloaded all publicly available metadata in Qiita and processed it into attribute-value pairs **(Figure 1A)**. In total, these data comprise 39,526,432 attribute-value pairs across 310,509 samples and 534 studies, encompassing both information describing the samples in a study and prep information that describes how the samples were processed in the wet lab. Preliminary analyses of these data indicate that many of the annotation issues we previously noted in SRA metadata are also prevalent in Qiita metadata. These include missing values and redundant attributes **(Figure 3A-B)**, addition of a large number of nonstandard attributes not recommended or required by Qiita **(Figure 3C)**, and typos and differences in capitalization **(Figure 3D)**. A more systematic analysis will be performed to determine which attributes will most benefit from attribute-value learning, but these preliminary data indicate that Qiita metadata quality can be substantially improved.

Our approach uses word embedding models trained on large text corpora to numerically capture contextual and semantic information of words by representing word similarity as a geometric distance in an n-dimensional space (4) (**Figure 1B)**. These word embeddings will first be used to cluster metadata categories with a high degree of similarity in the embedding space (5). The category embeddings can reveal a lack of normalization in category naming, mostly in the form of small deviations in spelling and punctuation (e.g., cell type and Cell type) and can highlight cases where values under the same attribute vary substantially in concept, or distinct attributes share similar values. After identifying problematic attribute-value pairs, we will remove them and retain the remaining attribute-value pairs to develop our NLP deep learning model. Further analysis of retained attribute values for target attributes will be performed in word embedding space to select the optimal values for classifier training.

We will train a bi-directional long-short term memory (bi-LSTM) recurrent neural network (**Figure 1C**) for metadata category classification, as bi-LSTMs have been shown to capture the sequential nature of text and the short and long term relationships between words (6). We hypothesize that a bi-LSTM model retrained on Qiita metadata can be used to automatically impute missing fields or correct non-standard values in microbiome data in Qiita or other microbiome datasets as well. To build and validate the classifier, we will split examples by study into training and test sets. We previously found that a publicly available Word2vec model trained on the entire PubMed, PMC and Wikipedia text corpora and included 5,443,656 word vectors, each with 200 features (7), paired with a bi-LSTM architecture, performed well on the metadata classification and prediction tasks.

The trained classifier will be used to locate metadata categories from longer, unstructured text using NER. Specifically, given a block of text **(Figure 1D)**, this model can predict which of a predefined set of metadata categories each n-gram of words in that text is most likely describing. Before metadata prediction, blocks of text are first split into sentences based on common sentence delimiters ‘;’’,’’.’. Whitespaces are then replaced with a single whitespace character, ‘ ’, sentences are tokenized using the python nltk package (v3.4.5), and empty tokens and stop words are removed. In each sentence, all possible n-grams of length 2 to 7 with at least 2 tokens in the Word2vec supported vocabulary, are assigned a score for each metadata category with the trained model. Each n-gram is then assigned to the metadata category with the highest score. Low confidence predictions, e.g. n-grams for which the difference in the highest and second highest scoring categories is less than or equal to 0.1, are removed. If multiple n-grams are classified to the same metadata category, the highest scoring n-gram of all overlapping n-grams is retained, with all others discarded **(Figure 1E)**. We may also leverage the named-entity extraction and normalization models for diseases, bacteria, and medications that co-Pi Hsu and colleagues developed(8-10). These models were designed for text mining from scientific documents and will be used only when publications are available.

Performance will be evaluated against a sample of metadata from Qiita with high quality annotations that can serve as “ground truth” values. To determine prediction accuracy, we compare the cosine similarity in the word-embedding space between corresponding predicted and ground truth values. If this value is greater than 0.7, or if the entire predicted value was a substring contained within the ground truth value (or vice-versa), we consider this a match **(Figure 1F)**.

*Expected outcomes, potential pitfalls and alternative strategies:* This aim will implement automated harmonization of key metadata categories in Qiita and is expected to improve the consistency and quality of annotations. If determining the target field “attributes” to apply attribute-value learning to is difficult due to the large amount of diversity in the sample type and preparation techniques between sub-disciplines in microbiome research, we will revisit the approach within datasets stratified based on coarse grained approach.

We will also investigate if other word embedding representations may improve the performance further. First, we will consider subword embeddings. Instead of representing each token as a vector in a high dimensional space, subword embeddings represent subwords, such as a prefix and suffix. Subword embeddings may be useful in biomedical domains where terminology is made up of compositions of subwords, e.g., `neuropathy' is from `neuro' and `-pathy'. Subwords also account for inflectional and American vs. British variations (e.g., necrotizing vs. necrotising) and are robust against typos. We will consider BPEmb and FastText (both described in (11)).

Second, contextualized word embedding models based on transformer models, such as ELMo (12) and BERT (13), and more recently, GPT3(14) and T5(15). These models are capable of representing words and sentences given their document-level context and have been shown to significantly improve performance for a wide range of NLP tasks. Since these word embedding models may encode the same word differently under different contexts, our hypothesis is that they may not be optimal for short phrases in metadata without a context but we would like to empirically verify if this is the case. Meanwhile, several BERT models pre-trained using large corpora of the biomedical literature are available in the public domain. We may test them to compare their effectiveness for metadata harmonization. We will consider BioBERT(16), a BERT model trained on Wikipedia, BookCorpus, PubMed, and PubMed Central, and SciBERT(17), which was trained on the scientific literature corpus from Semantic Scholar (<https://www.semanticscholar.org>).

Aim 2: Create a webtool frontend to support user metadata entry when adding datasets to Qiita

The goal of this aim is to streamline and normalize user metadata input at the time of dataset upload. To do this, we will develop a webtool that will populate the metadata attributes for each sample based on a block of text. We anticipate that abstracts from publications will provide the optimal source of information for this task. Thus, we will first adapt the approach in Aim 1 to predict MIxS and human phenotype attributes from abstracts. This would entail revisiting prioritization of the best assigned category from larger text blocks.

Users would have the opportunity to check the correctness of predicted attribute labels and change them when appropriate. Labels that could not be inferred from the provided text would be highlighted to attract the user’s attention. The webtool would further map the user input in each field into word embedding space and evaluate similarity to other attributes populating that field. The tool would propose the most similar values as possible alternative attribute values, ranked by similarity and frequency in the database. The user would then have the opportunity to select an appropriate value from this list, or retain their own input. The service may be designed to generate an Excel spreadsheet with metadata headers and values pre-filled or an online Google Spreadsheet, similar to the design of Qiimp (https://qiita.ucsd.edu/iframe/?iframe=qiimp). We expect to work with the Qiita development team to integrate this web tool into the Qiita data upload workflow.

*Expected outcomes, potential pitfalls and alternative strategies:* This aim will improve the Qiita metadata submission process on the user-side and the repository-side. This improved metadata in turn will improve future meta-analysis with Qiime2, Qiita and other microbiome analysis toolkits.

Aim 3: Perform a metaanalysis of IBD using studies retrieved post-metadata harmonization

In order to evaluate the benefit of Qiita metadata harmonization for metaanalysis, initially, we propose to revisit the study performed by Gonzalez *et al.* on Qiita (2) using our improved metadata and Qiita’s analysis platform. If harmonization was effective, we hypothesize we will be able to reproduce reported biological associations, and potentially detect novel ones, particularly if metadata harmonization allows integration of additional datasets or improves combined statistical analysis by reducing noise in user-defined values of important covariates (e.g. sex, age, diet and ethnicity).

In our meta-analysis, we plan on following the microbiome analysis pipeline outlined in (2). As was done by Gonazles *et al.*, we will use the Qiita web-interface to perform an principal coordinates analysis (PCoA) with unweighted unifrac (18) as a distance metric, followed by calculating distances of samples to the “healthy plane” using the methodology outlined in (19). The meta-analysis originally reported in Gonzalez et al. integrated three studies in 2018. We anticipate that more studies should be available to be included two years later, and the improved metadata should allow us to rapidly identify these studies. Further analysis steps, such as alpha diversity analysis and differential abundance testing, can also be integrated via Qiita or Qiime2 (20). In addition to reproducing published findings, we will collaborate with CMI faculty member Trey Ideker to build species co-abundance networks, and novel co-abundance networks of enzymatic functions inferred from the genes present in the species detected in healthy controls versus IBD cases. Analysis of these networks will provide additional insights into the biological mechanisms underlying IBD.

*Expected outcomes, potential pitfalls and alternative strategies:* This aim will produce a report summarizing a concrete example of the benefit of metadata harmonization for meta-analysis with Qiita. If additional relevant studies cannot be incorporated post-harmonization, we will identify the barriers to doing so and evaluate whether our approach in Aim 1 can be improved to address the issue. If we fail to see expected patterns or are unable to recapitulate or find new biological insights from this meta-analysis, we will identify and remove datasets with low quality annotations.

**Reporting Timeline:**

|  |  |
| --- | --- |
| Months 1-6 | Parallel development of Qiita NLP harmonization tool and webtool for metadata entry, at the end of 6 months, submit proposed updates to Qiita metadata |
| Months 7-12 | Perform meta-analysis on IBD, and beta testing of the webtool |
| Month 12 | Web tool complete and functional, complete metadata analysis and prepare manuscript |

**Deliverables:**

* A deep learning tool trained specifically for Qiita metadata harmonization
* A web tool frontend to support metadata entry for new studies added to Qiita
* An updated meta-analysis of IBD post data-harmonization describing the benefit for discovery

**Potential Commercial Impact:** This proposal will develop a deep learning framework with a frontend web tool for automated metadata entry and cleaning. The approach itself can be adapted for various databases. Meta-data harmonization will increase the ease with which the Qiita database can be searched for datasets to combine for meta-analysis, leading to better powered studies and accelerating new findings in microbiome analysis.

**Budget:** We are requesting the full $200,000 to support this effort:

Direct costs: $135,217 is requested to cover personal effort and expenses.

*Personnel:* $56,133 to support graduate student Adam Klie in Carter lab, $11,700 for an hourly rate student researcher in Hsu lab, $24,121 to support Dr. Carter (Fringe benefit $5,217included), $23,150 to support Hsu (Fringe benefit $7,052 included) and $8,855 to support Dr. Ideker (Fringe benefit $1,915 included) to perform the work described in the proposal. *Expenses:* $10,000 for publication costs and travel to present at meetings.

Indirect costs: At an IDC rate of 58%, indirect costs are $66,707.

**References:**

1. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PS, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methé BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spo A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glöckner FO. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29(5):415-20. doi: 10.1038/nbt.1823. PubMed PMID: 21552244; PMCID: PMC3367316.

2. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods. 2018;15(10):796-8. Epub 2018/10/01. doi: 10.1038/s41592-018-0141-9. PubMed PMID: 30275573; PMCID: PMC6235622.

3. Klie A, Tsui BY, Mollah S, Skola D, Dow M, Hsu C-N, Carter H. Increasing metadata coverage of SRA BioSample entries using deep learning based Named Entity Recognition. bioRxiv. 2020:414136. doi: 10.1101/414136.

4. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.

5. Gonçalves RS, Kamdar MR, Musen MA, editors. Aligning biomedical metadata with ontologies using clustering and embeddings. European Semantic Web Conference; 2019: Springer.

6. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 2005;18(5-6):602-10. doi: 10.1016/j.neunet.2005.06.042. PubMed PMID: 16112549.

7. Chiu B, Crichton G, Korhonen A, Pyysalo S, editors. How to train good word embeddings for biomedical NLP. Proceedings of the 15th workshop on biomedical natural language processing; 2016.

8. Wright D. NormCo: Deep disease normalization for biomedical knowledge base construction: UC San Diego; 2019.

9. Hogan W, Mehta R, Vazquez-Baeza Y, Katsis Y, Kim H-C, Hsu C-N. Normalization of Predominant and Long-tail Bacterial Entities with a Hybrid CNN-LSTM and Knowledge-Driven Model.

10. Weaver CKH, William P. Mehta, Raghav Yang, Xuhan Katsis, Yannis Vazquez- Baeza, Yoshiki, Swafford A, Kim H-C, Hsu C-N. Creating a knowledge base of diseases and microbes associated with covid-19. In preparation.

11. Heinzerling B. BPEmb: Pre-trained Subword Embeddings in 275 Languages (LREC 2018)2019.

12. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv preprint arXiv:180205365. 2018.

13. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

14. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners. arXiv preprint arXiv:200514165. 2020.

15. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:191010683. 2019.

16. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.

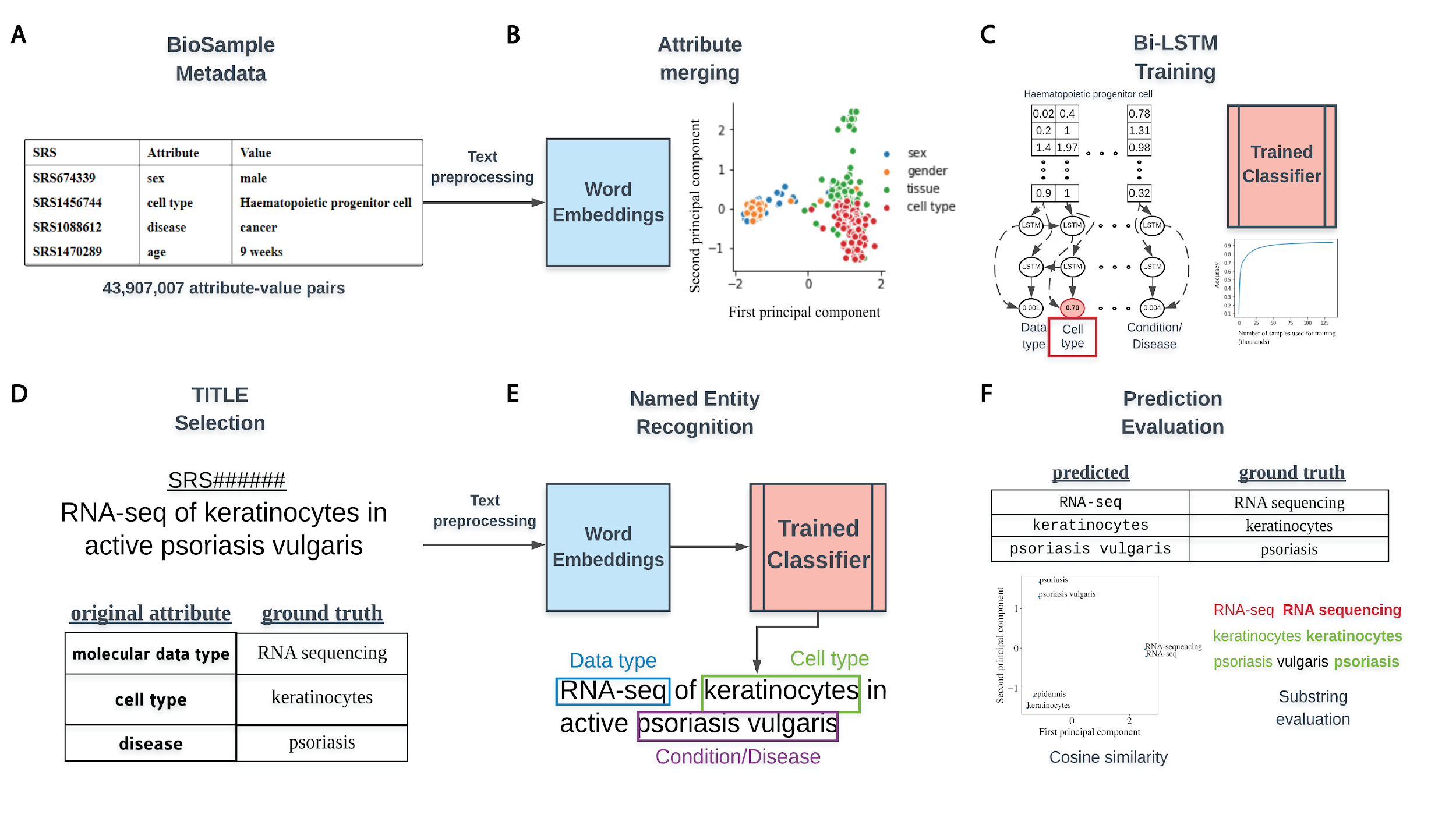
17. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:190310676. 2019.

18. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12):8228-35. doi: 10.1128/AEM.71.12.8228-8235.2005. PubMed PMID: 16332807; PMCID: PMC1317376.

19. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, D'Amato M, Bonfiglio F, McDonald D, Gonzalez A, McClure EE, Dunklebarger MF, Knight R, Jansson JK. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol. 2017;2:17004. Epub 2017/02/13. doi: 10.1038/nmicrobiol.2017.4. PubMed PMID: 28191884; PMCID: PMC5319707.

20. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852-7. doi: 10.1038/s41587-019-0209-9. PubMed PMID: 31341288; PMCID: PMC7015180.

**Figures:**



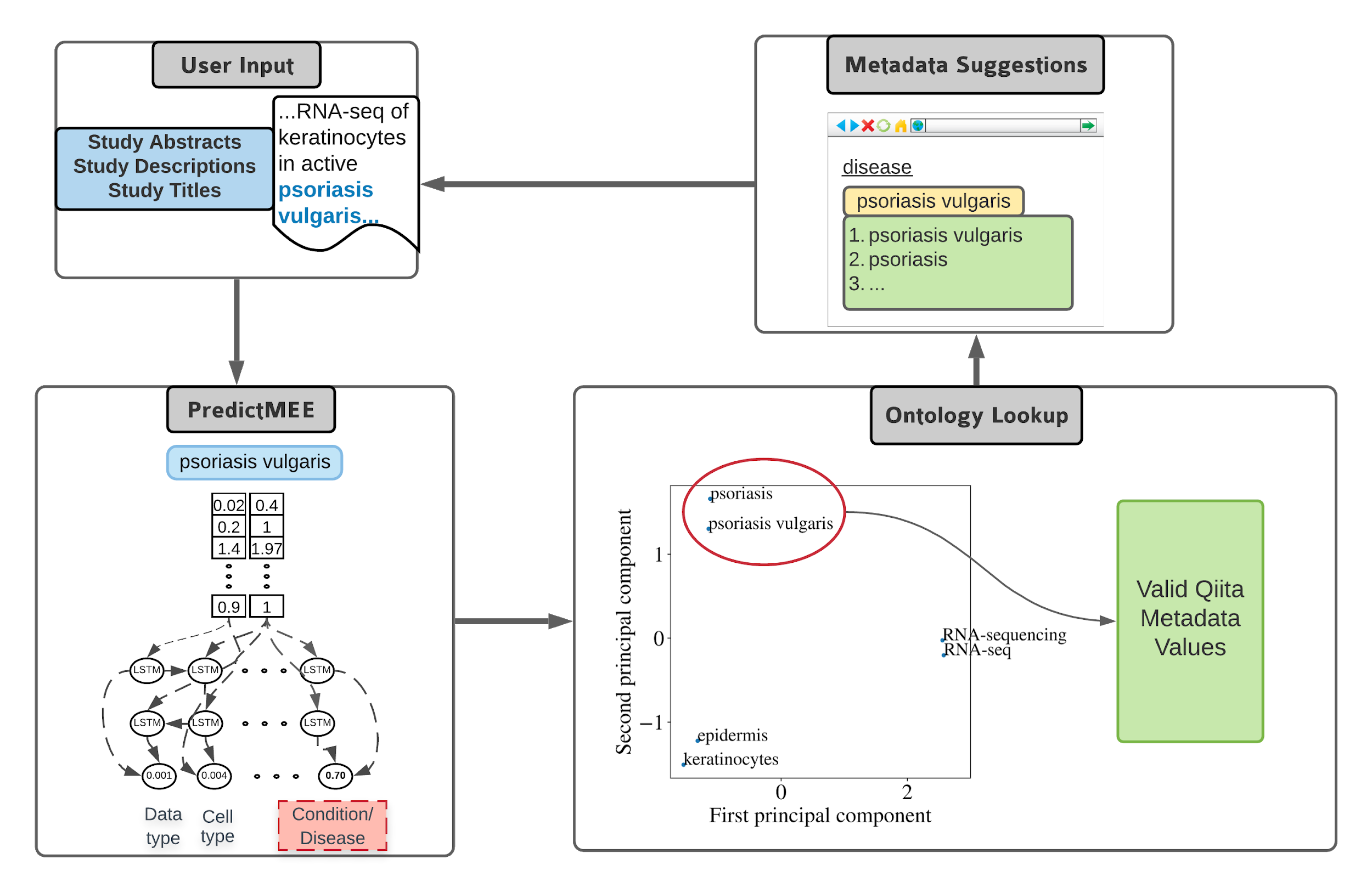
**Figure 1. Existing PredictMEE Workflow** Overview of PredictMEE metadata prediction workflow as previously applied to SRA. **(A)** Example attribute-value pairs in SRA BioSample. **(B)** Clustering of attribute values in word embedding space allows identification and merging of attributes describing similar concepts. **(C)** A subset of attribute-value pairs was split into a training and test set and a bi-LSTM classifier was trained to identify 11 metadata categories. **(D)** TITLEs were selected as the text block for NER using the trained model. An example TITLE with associated ground truth labels is shown. **(E)** These TITLEs were preprocessed into n-grams and fed into the trained classifier after word embeddingto generate metadata predictions for the 11 categories. **(F)** Comparisons to ground truth metadata were done using substring matching and cosine similarity in the word embedding space.

## 

**Figure 2.** **Previous performance of bi-LSTM in metadata category classification on SRA. (A)** Accuracy, precision, recall, F1 score, and average AUROC calculated for PredictMEE model trained on SRA for all 11 categories combined. **(B)** Accuracy of model classification on training set (y-axis) plotted against the number of training examples input (in thousands). **(C)** Percentage of each category correctly classified, shown as a heatmap, with predicted values on the x-axis and ground truth labels on the y-axis. **(D)** Receiver operating characteristic (ROC) curves for each category along with the average over all test set examples (micro average).

****

**Figure 3. Human-gut Qiita metadata shows substantial heterogeneity** **(A)** Qiita coverage in human host samples for a set of manually selected metadata attributes. Attributes were chosen from Qiita documentation (2) and MIxS (v5) specification (1). **(B)** Qiita coverage in human-gut samples for 46 Qiimp-suggested attributes for the “human-gut” environmental package. **(C)** Overlap of the 46 Qiimp-suggested attributes with all attributes present in Qiita metadata for human-gut samples. **(D)** Examples of normalization typos for metadata of human samples in Qiita. Cosine similarity indicates the semantic similarity between the attributes in the word embedding space.

****

**Figure 4. Web-based tool schematic for automating metadata suggestions and fill-in** Proposed workflow for automated metadata prediction using PredictMEE. Submitted abstracts, titles descriptions or other free-text can be processed and fed through our existing NLP framework to generate predicted metadata. The closest ontology terms to the predicted metadata in the word embedding space can be used to generate a list of potential metadata for users to select. Low confidence and missing predictions will be highlighted for the user to manually fill-in and these metadata can then be integrated with Qiimp to autofill excel or Google spreadsheets (not shown here).

OMB No. 0925-0001 and 0925-0002 (Rev. 09/17 Approved Through 03/31/2020)

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.

Follow this format for each person. DO NOT EXCEED FIVE PAGES.

|  |
| --- |
| NAME: Carter, Hannah Kathryn |
| eRA COMMONS USER NAME (credential, e.g., agency login): HKCARTER |
| POSITION TITLE: Assistant Professor |

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

|  |  |  |  |
| --- | --- | --- | --- |
| INSTITUTION AND LOCATION | DEGREE (if applicable) | END DATE MM/YYYY | FIELD OF STUDY |
| University of Louisville, Louisville, Kentucky | MENG | 12/2004 | Electrical and Computer Engineering |
| Johns Hopkins University, Baltimore | PHD | 05/2012 | Biomedical Engineering |
| University of California, San Diego, California | Postdoctoral Fellow | 10/2013 | Postdoctoral training in cancer systems biology under Trey Ideker |

### A. Personal Statement

My qualifications to serve as co-PI for this proposal lie with my extensive training in computational data analysis, experience analyzing and integrating diverse datasets and datatypes to address unanswered biomedical questions, and history of success leading multi-disciplinary research teams. I trained in cancer genomics with Rachel Karchin and Bert Vogelstein at Johns Hopkins University from 2006 to 2012, then joined the Ideker Lab at UCSD as a post-doc to train in computational systems biology. My interests are focused on genomic analysis to understand genetic mechanisms underlying disease. My research addresses three major areas of need for precision medicine: 1) identifying inherited risk factors that contribute to disease risk 2) developing bioinformatic tools to aid interpretation genetic variation in the context of disease risk and progression and 3) translating molecular measurement data to therapeutic opportunities. I have over a decade’s experience processing and critically analyzing large multi-omic datasets including The Cancer Genome Atlas, GTEx, ENCODE, etc. The machine learning tools I previously developed to identify functional genetic variants (Carter et al 2009, Carter et al 2013) are now widely used by the cancer research and variant interpretation communities.

Our recent work most relevant to this proposal is twofold: 1) In a collaborative effort with the Hsu lab, we recently developed the PredictMEE deep learning tool that uses natural language processing approaches to address metadata issues with large biomedical databases. This work is currently under peer review. Secondly, my lab’s recent research efforts have focused on evaluating interactions between the inherited genome and the evolving tumor genome. After an early screen suggested that common variation could associate with the specific driver mutations observed in tumors (Carter et al 2017), we focused on the inherited immune system as a likely source of strong selective pressure during tumor evolution. We devised a score that maps the highly diverse Major Histocompatibility Complex class I genotype across individuals with cancer into a shared space of driver mutation derived neoantigens, enabling study of polymorphic differences in this immune molecule in the context of tumor evolution at scale (Marty et al 2017). A second manuscript (Marty et al 2018) extended this analysis to MHC class II and the CD4+ T cell response in shaping the tumor genome. We also showed that somatic mutations affecting presentation by MHC class I are associated with differences in tumor mutation burden, neoantigen burden and characteristics of the tumor-immune microenvironment in microsatellite stable tumors (Castro et al 2019). Our experience co-analyzing the contents of co-existing genomes, and experience with the human immune system provide us with unique perspectives that this proposal will bring to microbiome studies.

For the current proposal, I will supervise bioinformatic tool development and data analysis by gradaute student Adam Klie. I will work closely with Dr. Hsu who has extensive expertise in deep learning and natural language processing to adapt our approaches for Qiita and with Dr. Ideker to design co-abundance network analyses for the metanalysis of IBD planned in Aim 3.

1. Klie A, Tsui BY, Mollah S, Skola D, Dow M, **Hsu C-N**, **Carter H**. Increasing metadata coverage of SRA BioSample entries using deep learning based Named Entity Recognition. bioRxiv. 2020:414136. doi: 10.1101/414136.
2. Marty Pyke R, Thompson WK, Salem RM, Font-Burgada J, Zanetti M, Carter H. Evolutionary Pressure against MHC Class II Binding Cancer Mutations. Cell. 2018 Oct 4;175(2):416-428.e13. PubMed PMID: [30245014](http://www.ncbi.nlm.nih.gov/pubmed/30245014/); PubMed Central PMCID: [PMC6482006](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6482006/).
3. Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB, de Prisco N, Ideker T, Hildebrand WH, Font-Burgada J, Carter H. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. Cell. 2017 Nov 30;171(6):1272-1283.e15. PubMed PMID: [29107334](http://www.ncbi.nlm.nih.gov/pubmed/29107334/); PubMed Central PMCID: [PMC5711564](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5711564/).
4. **Carter H**, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, Wu X, DeBoever C, Van Nostrand EL, Song Y, Wheeler E, Kreisberg JF, Lippman SM, Yeo GW, Gutkind JS, **Ideker T**. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. Cancer Discov. 2017 Apr;7(4):410-423. PubMed PMID: [28188128](http://www.ncbi.nlm.nih.gov/pubmed/28188128/); PubMed Central PMCID: [PMC5460679](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5460679/).

### B. Positions and Honors

Positions and Employment

|  |  |
| --- | --- |
| 2014 – 2020  2020 - | Assistant Professor, University of California, San Diego, La Jolla, CA  Associate Professor, University of California, San Diego, La Jolla, CA |

Other Experience and Professional Memberships

|  |  |
| --- | --- |
| 2000 - 2004 | Member, Society of Women Engineers (SWE) |
| 2002 - 2004 | Member, Institute for Electrical and Electronics Engineers (IEEE) |
| 2008 - | Associate Member, American Association for Cancer Research (AACR) |
| 2008 - | Member, International Society for Computational Biology (ICSB) |
| 2011 - | Member, American Association for the Advancement of Science (AAAS) |

Honors

|  |  |
| --- | --- |
| 1999 | Trustees’ Scholarship, University of Louisville |
| 2000 | Overseer's Scholar, University of Louisville |
| 2003 | Theobald Scholarship, University of Louisville |
| 2004 | Electrical and Computer Engineering Department Alumni Award, University of Louisville |
| 2005 | Samuel T. Fife Outstanding Graduate Award, University of Louisville |
| 2008 | National Defense Science and Engineering Graduate Fellowship, Department of Defense |
| 2012 | Siebel Scholar, Siebel Foundation |
| 2013 | NIH Director’s Early Independence Award, NIH |
| 2015 | Alumni Outstanding Recent Graduate Award, Johns Hopkins University |
| 2017 | Azrieli Global Scholar, Canadian Institute For Advanced Research (CIFAR) |
| 2019 | Emerging Leader and Jaime Wyatt Miller Fellow, The Mark Foundation |

### C. Contribution to Science

1. Delineating the molecular mechanisms that drive tumorigenesis remains a central goal in cancer research. The availability of large tumor datasets with multiple 'omics layers presents great opportunities for better characterizing the mechanisms underlying various neoplastic behaviors of cancer cells. However the sheer complexity of biological systems is challenging to capture with simple data mining approaches. A central goal of my research program is to devise new methods and strategies to improve mining of biological mechanisms from large cancer 'omics datasets. The methods developed so far have aimed at better identification of subgroups of patients with common molecular phenotypes to determine biological pathways specific to disease subgroup, finding interactions between somatic molecular changes that have important prognostic implications for patients and building more detailed models to better capture the specific biological pathways perturbed by somatic mutations.
   1. Engin HB, Kreisberg JF, Carter H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. PLoS One. 2016;11(4):e0152929. PubMed PMID: [27043210](http://www.ncbi.nlm.nih.gov/pubmed/27043210/); PubMed Central PMCID: [PMC4820104](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4820104/).
   2. Engin HB, Hofree M, Carter H. Identifying mutation specific cancer pathways using a structurally resolved protein interaction network. Pac Symp Biocomput. 2015;PubMed PMID: [25592571](http://www.ncbi.nlm.nih.gov/pubmed/25592571/); PubMed Central PMCID: [PMC4299875](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299875/).
   3. Gross AM, Orosco RK, Shen JP, Egloff AM, Carter H, Hofree M, Choueiri M, Coffey CS, Lippman SM, Hayes DN, Cohen EE, Grandis JR, Nguyen QT, Ideker T. Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss. Nat Genet. 2014 Sep;46(9):939-43. PubMed PMID: [25086664](http://www.ncbi.nlm.nih.gov/pubmed/25086664/); PubMed Central PMCID: [PMC4146706](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4146706/).
   4. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013 Nov;10(11):1108-15. PubMed PMID: [24037242](http://www.ncbi.nlm.nih.gov/pubmed/24037242/); PubMed Central PMCID: [PMC3866081](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3866081/).
2. With the first large-scale tumor exome sequencing studies it quickly became apparent that new strategies would be needed to prioritize the large number of protein altering variants detected. In order to study biological mechanisms underlying tumorigenesis, neutral 'passenger' mutations needed to be distinguished from causal 'driver' mutations. To this end, I developed the first tumor-specific method for prioritizing somatic missense mutations. This method, called CHASM, is the first method to explicitly model the mutation spectra of the tumor type in question when prioritizing somatic mutations. This is particularly important since the processes by which mutations arise in a tumor are distinct from processes by which they are fixed in the general population, and the mechanisms that generate the mutations are often cancer specific. CHASM has been widely adopted by the cancer research community and has been integrated into the Firehose pipeline used to process all of The Cancer Genome Atlas Data.
   1. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008 Sep 26;321(5897):1801-6. PubMed PMID: [18772397](http://www.ncbi.nlm.nih.gov/pubmed/18772397/); PubMed Central PMCID: [PMC2848990](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2848990/).
   2. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA Jr, Hartigan J, Smith DR, Strausberg RL, Marie SK, Shinjo SM, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008 Sep 26;321(5897):1807-12. PubMed PMID: [18772396](http://www.ncbi.nlm.nih.gov/pubmed/18772396/); PubMed Central PMCID: [PMC2820389](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2820389/).
   3. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009 Aug 15;69(16):6660-7. PubMed PMID: [19654296](http://www.ncbi.nlm.nih.gov/pubmed/19654296/); PubMed Central PMCID: [PMC2763410](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2763410/).
   4. Carter H, Samayoa J, Hruban RH, Karchin R. Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM). Cancer Biol Ther. 2010 Sep 15;10(6):582-7. PubMed PMID: [20581473](http://www.ncbi.nlm.nih.gov/pubmed/20581473/); PubMed Central PMCID: [PMC3040948](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3040948/).
3. The causal variation underlying many Mendelian disorders and most common genetic diseases have yet to be elucidated. Next generation sequencing methodologies have enabled new strategies for detecting the underlying disease causing genes and mutations. I have previously developed bioinformatics methods aimed at better using sequencing data to track down the responsible regions of the genome as well as probabilistic methods to predict phenotype from genotype. By prioritizing genes based on bioinformatic mutation scores and analyzing gene ranks across patients, it was possible to identify Mendelian disease genes using only a small number of genomes (Carter et al 2013). To further support identification of disease genes, I worked on a burden testing approach that would enable mutation burden to be evaluated at the sub-gene level, allowing detection of protein domains biased towards functional mutations (Chen et al 2013). I also led the development of a method to predict a broad spectrum of phenotypes from whole genome sequencing data for two consecutive meetings of CAGI (Critical Assessment of Genome Interpretation). This method was later formalized and published (Chen et al 2014).
   1. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3. PubMed PMID: [23819870](http://www.ncbi.nlm.nih.gov/pubmed/23819870/); PubMed Central PMCID: [PMC3665549](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3665549/).
   2. Chen YC, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, Zandi PP, McCombie WR, Potash JB, Karchin R. A hybrid likelihood model for sequence-based disease association studies. PLoS Genet. 2013;9(1):e1003224. PubMed PMID: [23358228](http://www.ncbi.nlm.nih.gov/pubmed/23358228/); PubMed Central PMCID: [PMC3554549](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3554549/).
   3. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics. 2013 Mar 1;29(5):647-8. PubMed PMID: [23325621](http://www.ncbi.nlm.nih.gov/pubmed/23325621/); PubMed Central PMCID: [PMC3582272](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3582272/).
   4. Chen YC, Douville C, Wang C, Niknafs N, Yeo G, Beleva-Guthrie V, Carter H, Stenson PD, Cooper DN, Li B, Mooney S, Karchin R. A probabilistic model to predict clinical phenotypic traits from genome sequencing. PLoS Comput Biol. 2014 Sep;10(9):e1003825. PubMed PMID: [25188385](http://www.ncbi.nlm.nih.gov/pubmed/25188385/); PubMed Central PMCID: [PMC4154636](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4154636/).
4. We have identified new inherited factors that influence the molecular characteristics in tumors. The role of the germline line in predisposing individuals to cancer is still poorly understood. We are developing strategies to determine germline factors that underlie tumor development including interactions between germline variation and somatic alterations that emerge during tumor development and other clinical covariates such as the site at which a tumor developed or age at the time of diagnosis. We recently reported new germline loci linked to increased somatic mutation rates of cancer genes via shared biological pathways. For example, a haplotype encoding the RNA binding protein RBFOX1, a regulator or exon inclusion during splicing, increased the incidence of SF3B1 mutations, known to alter splice site recognition (Carter et al 2017). We have also demonstrated that individual genotype at the HLA locus encoding MHC-I determines the set of cancer causing mutations that can be exposed to an individual's immune system via the antigen presentation pathway. We found that mutations were more likely to be observed in the tumors of individual's who's MHC-I genotype precluded those mutations from being effectively presented to T-cells (Marty et al 2017), with even stronger effects being observed for MHC-II genotype (Marty Pyke et al 2018).
   1. Marty Pyke R, Thompson WK, Salem RM, Font-Burgada J, Zanetti M, Carter H. Evolutionary Pressure against MHC Class II Binding Cancer Mutations. Cell. 2018 Oct 4;175(2):416-428.e13. PubMed PMID: [30245014](http://www.ncbi.nlm.nih.gov/pubmed/30245014/); PubMed Central PMCID: [PMC6482006](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6482006/).
   2. Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB, de Prisco N, Ideker T, Hildebrand WH, Font-Burgada J, Carter H. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. Cell. 2017 Nov 30;171(6):1272-1283.e15. PubMed PMID: [29107334](http://www.ncbi.nlm.nih.gov/pubmed/29107334/); PubMed Central PMCID: [PMC5711564](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5711564/).
   3. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, Wu X, DeBoever C, Van Nostrand EL, Song Y, Wheeler E, Kreisberg JF, Lippman SM, Yeo GW, Gutkind JS, Ideker T. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. Cancer Discov. 2017 Apr;7(4):410-423. PubMed PMID: [28188128](http://www.ncbi.nlm.nih.gov/pubmed/28188128/); PubMed Central PMCID: [PMC5460679](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5460679/).
   4. Levy E, Marty R, Gárate Calderón V, Woo B, Dow M, Armisen R, Carter H, Harismendy O. Immune DNA signature of T-cell infiltration in breast tumor exomes. Sci Rep. 2016 Jul 25;6:30064. PubMed PMID: [27452728](http://www.ncbi.nlm.nih.gov/pubmed/27452728/); PubMed Central PMCID: [PMC4958917](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4958917/).

### D. Additional Information: Research Support and/or Scholastic Performance

Ongoing Research Support

1R01CA220009-01, NCI

Zanetti (PI)

08/17/17-07/31/22

(PQ3) Disruption of immune surveillance by aneuploidy and aberrant MHCII expression

This project investigates the relationship between aneuploidy, antigen presentation and response to checkpoint inhibitors.

Role: CPI

Emerging Leader Award, The Mark Foundation

Carter, Hannah Kathryn (PI)

03/15/19-03/14/22

Enabling MHC genotype-informed risk prediction, cancer prevention and precision immunotherapy

This project will develop MHC genotype informed models to predict cancer susceptibility and potential to respond to immunotherapy therapy

Role: PI

20191857, Harry J Lloyd Charitable Trust

Carter, Hannah Kathryn (PI)

06/01/19-05/31/20

Characterizing germline and somatic determinants of anti-tumor immunity in melanoma

This project aims to optimize precision immunotherapy for melanoma

Role: PI

Completed Research Support

P50GM085764-07, NIGMS

Hasty and Ideker (PI)

07/01/08-05/31/19

San Diego Center for Systems Biology: From Maps to Models

Center grant funding a multi-PI systems biology effort

Role: Co-Investigator

DP5 OD017937-01

Carter, Hannah Kathryn (PI)

09/24/13-08/31/18

Network approaches to identify cancer drivers from high-dimensional tumor data

Role: PI

FL-000655, CIFAR Canadian Institute for Advanced Research

Carter, Hannah Kathryn (PI)

07/01/17-06/30/19

Genetic Networks Program: Systematic screening to detect germline-somatic interactions in cancer

Role: PI

OMB No. 0925-0001 and 0925-0002 (Rev. 03/2020 Approved Through 02/28/2023)

**BIOGRAPHICAL SKETCH**

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

NAME: Chun-Nan HSU, PhD

eRA COMMONS USER NAME (credential, e.g., agency login): chunnanhsu

POSITION TITLE: Project Scientist, UC San Diego; WOS Computer Scientist VASDHS

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

| INSTITUTION AND LOCATION | DEGREE  *(if applicable)* | Completion Date  MM/YYYY | FIELD OF STUDY |
| --- | --- | --- | --- |
| National Chiao Tung University | BS | 06/1988 | Computer Engineering |
| University of Southern California | MS | 05/1992 | Computer Science |
| University of Southern California | PhD | 12/1996 | Computer Science |
|  |  |  |  |

**A. Personal Statement**  
After I earned my PhD in Computer Science, I was appointed Assistant Professor in Computer Science & Engineering at Arizona State University before I joined Academia Sinica, Taiwan (1998-2014) and was promoted to a tenured position in 2008. I led the Informatics group of the Advanced Bioinformatics Core Service of National Research Program in Genomic Medicine (2006-2011). My team developed widely used software tools for biomedical sciences, including tools for celluar image analysis and SNP prioritization. I was elected as the 8th President of the Taiwanese Association for Artificial Intelligence (2009-2011). Before joining UCSD, I was a Computer Scientist at the Information Sciences Institute, University of Southern California (2009-2013), where I was the main author of the grant proposal and key-investigator of the NHLBI PFINDR project. My research has also contributed to the basic research of machine learning. I developed new machine learning algorithms to scale up training of conditional random fields, Bayesian networks, support vector machines, and Deep Convolutional Neural Networks using big data in 2009 [a]. At UCSD, I have been leading teaching and research in biomedical natural language processing (NLP) and text mining. I served as PI for an NHGRI-funded bioNLP project aiming at automatically curating a catalog of genome-wide association studies (GWAS), and more recently a DoD-funded project on Deep Learning for military medicine. I have been teaching the course “Introduction to Biomedical Natural Language Processing” for both graduate (engineering and biology) and medical students. I have published more than 100 highly cited peer-reviewed research articles in the fields of machine learning, data mining, biomedical informatics, and AI for healthcare (e.g., [b]). I was awarded Senior Member of Association of Computing Machinery (ACM) in 2011 and the IBM Faculty Award in 2012 for the distinguished contributions to biomedical text mining.

The proposed project will be based on my recent work with Dr. Carter. I am confident that the proposed approach and the excellent team will deliver a solution that unleashes the potential of microbiome data repositories to facilitate meta-analysis and contributes to new findings.

1. Hsu CN, Chang Y, Huang H, Lee Y. Advances in Neural Information Processing Systems 22 (NIPS 2009). Red Hook, NY: Curran Associates, Inc.; 2009. Periodic Step Size Adaptation for Single Pass On-line Learning. 763-771p.
2. Liang H, et al., Hsu CN, Carter H, Zhu L, Zhang K, and Xia H. Evaluation and accurate diagnoses of

pediatric diseases using artificial intelligence. Nature Medicine. 2019 Feb 11: p1. PMID: 30742121.

1. Rodney A. Gabriel, Tsung-Ting Kuo, Julian McAuley, and Chun-Nan Hsu. Identifying and characterizing highly similar notes in big clinical note datasets. Journal of biomedical informatics 82: 63-69, 2018.
2. Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. NormCo: Deep Disease

Normalization for Biomedical Knowledge Base Construction. The Conference of Automated Construction of Knowledge Bases (AKBC), May 21, 2019. (Best application paper award) <https://openreview.net/pdf?id=BJerQWcp6Q>

**B. Positions and Honors**

**Positions**

|  |  |  |  |
| --- | --- | --- | --- |
| 1996-1998 | Arizona State Univ. Dept Computer Sci. | Tempe, AZ | Assistant Professor |
| 1998-2000 | Academia Sinica, Inst. Information Sci. | Taipei, Taiwan | Assistant Research Fellow |
| 1999-2000 | Natl’ Chiao Tung Univ. Dept Comp Eng. | Hsinchu, Taiwan | Adjunct Assistant Professor |
| 2000-2005 | Deepspot Inc. | Taipei, Taiwan | Co-founder & Scientific Consultant |
| 2002-2008 | Academia Sinica, Inst. Information Sci. | Taipei, Taiwan | Associate Research Fellow |
| 2008-2014 | Acaedmia Sinica, Inst. Information Sci. | Taipei, Taiwan | Research Fellow |
| 2009-2011 | Taiwanese Assocation for AI | Taipei, Taiwan | President |
| 2009-2013 | University of Southern California | Los Angeles, CA | Computer Scientist |
| 2013- | UC San Diego, School of Medicine | La Jolla, CA | Adjt Asso. Prof. & Project Scientist |

**Honors**

|  |  |
| --- | --- |
| 1992 | Academic Achievement Award, University of Southern California |
| 1994 | Best Leadership Award, University of Southern California |
| 2002 | PhD Dissertation Advisor Award, Institute of Information and Computing Machinery, Taiwan |
| 2007 | 2nd Place Winner, BioCreative 2 |
| 2008 | Best Newcomer Award, Pascal Large Scale Learning Challenge |
| 2010  2011 | 1st Place Winner, BioCreative 3  Senior Member in the Association of Computing Machinery (ACM). (The Senior Member  Grade recognizes those senior ACM members who have demonstrated performance that sets them apart from their peers.) |
| 2012  2015  2016  2019 | IBM Faculty Award, IBM  Outstanding abstract award winner. Ed Quigley Clinical Research Symposium, Rady  Children’s Hospital, San Diego, CA (with Juan D. Chapparo et al.)  The Office of the National Coordinator for Health Information Technology (ONC) Use of Blockchain in Health IT and Health-related Research Challenge, Winner, UCSD Team. Best Application Paper Award. Conference of Automatic Knowledge Base Construction  (AKBC 2019) (as the corresponding author). |

**C. Contributions to Science**

**1. Biomedical Natural Language Processing**

I have contributed to the field of biomedical natural language processing (NLP) and text mining. My gene mention tagger achieved one of the best scores in the BioCreative 2 Challenge Evaluation in 2007. An improved version called AIIAGMT, applying an ensemble of conditional random field models that parse an input sentence in both forward and backward directions, was released in 2008 [a]. AIIAGMT not only achieved one of the best performances in gene mention tagging, the innovative idea of combining bi- directional parsing is now considered a standard practice in NLP. We implemented the algorithm into a free Web-based service at AIIAGMT, which was considered as the tool of choice for gene mention tagging. Our heavy users include NCBI (Z. Lu’s lab), Mouse Genome Initiative (J. Blake’s lab), University of Delaware (C. Wu’s Uniprot lab), and Carnegie Mellon University (Ziv’s lab). In September 2010, my team again ranked No.1 in four out of nine evaluation criteria among 14 teams in the gene normalization task of the BioCreative

3 Challenge Evaluation [b]. Some of the recent projects that I am involved include development of highly effective deep neural networks to diagnose a broad range of pediatric diseases [c] and to normalize disease

mentions in large collections of scientific publications [d].I received an IBM Faculty Award for my contributions in Biomedical Text Mining in 2012.

a. Hsu CN, Chang YM, Kuo CJ, Lin YS, Huang HS, Chung IF. Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics*. 2008 Jul 1;24(13):i286-94. PubMed PMID:[18586726](http://www.ncbi.nlm.nih.gov/pubmed/18586726/); PubMed Central PMCID: [PMC2718659](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2718659/). (BioCreative 2 Top Performer)

b. Kuo CJ, Ling MH, Hsu CN. Soft tagging of overlapping high confidence gene mention variants for cross-species full-text gene normalization. *BMC Bioinformatics*. 2011 Oct 3;12 Suppl 8:S6. PubMed PMID: [22152021](http://www.ncbi.nlm.nih.gov/pubmed/22152021/); PubMed Central PMCID: [PMC3269941](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3269941/). (BioCreative 3 Top Performer)

c. Liang H, et al., Hsu CN, Carter H, Zhu L, Zhang K, and Xia H. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine*. 2019 Feb 11: p1. PMID: [30742121](http://www.ncbi.nlm.nih.gov/pubmed/30742121/).

d. Wright D, Mehta R, Katsis Y, and Hsu CN. NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction. Proceedings of the Conference of Automated Construction of Knowledge Bases (AKBC), May 21, 2019. 15 pages. (Best application paper award) <https://openreview.net/pdf?id=BJerQWcp6Q>

e. Badal VD, Wright D, Katsis Y, Kim HC, Swafford AD, Knight R, Hsu CN. Challenges in the construction of knowledge bases for human microbiome-disease associations. *Microbiome*. 2019 Sep 5;7(1):129. doi: 10.1186/s40168-019-0742 2. Review. PubMed PMID: [31488215](https://www.ncbi.nlm.nih.gov/pubmed/31488215/); PubMed Central PMCID: [PMC6728997](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6728997/).

**2. Large-scale Machine Learning**

I have contributed in laying the foundations of applying machine learning to large data sets, what recently known as the "Big Data." To accelerate the training of various machine-learning algorithms with large high- dimensional training corpora in biological text mining, I developed an extrapolation method called the “Triple Jump” method [b], which is applicable to training algorithms for conditional random fields, Bayesian networks, support vector machines, and deep convolutional neural networks. With the method, my post-doc won the “best newcomer award” in the Large-Scale Learning challenge in the International Conference on Machine Learning (ICML) 2008. Then I extended our method to on-line learning and developed a new algorithm called Periodic Step-size Adaptation. We showed that PSA could converge to an empirical optimum in a single pass through the training examples. Single-pass online learning is interesting because that will minimize disk I/O when the size of training examples is too large to fit in a computer's memory [a,b], which is common in nature language processing applications. I was therefore invited to serve as a guest-editor for a journal special issue on the applications of large-scale machine learning in 2010 [d].

a. Hsu CN, Chang Y, Huang H, Lee Y. Advances in Neural Information Processing Systems 22 (NIPS

2009). Red Hook, NY: Curran Associates, Inc.; 2009. Periodic Step Size Adaptation for Single Pass

On-line Learning763-771p.

b. Huang H, Yang B, Chang Y, Hsu CN. Global and componentwise extrapolations for accelerating training of Bayesian networks and conditional random fields. Data mining and knowledge discovery.

2009 March 06; 19(1):58-94.

c. Hsu CN, Huang H, Chang Y, Lee Y. Periodic step-size adaptation in second-order gradient descent for single-pass on-line structured learning. Machine Learning. 2009 December 01; 77(2-3):195-224.

d. Hsu CN. Introduction to special issue on large-scale machine learning. ACM transactions on intelligent systems and technology. 2011 April; 2(3):25.

**3. Integration of Biomedical Data on the Web for Precision Medicine**

In 2004, I developed and released a software package called “Agent Toolbox” [a,b] that allows users to quickly produce and maintain a large number of “Web wrapper agents” without programming. A Web wrapper agent is a script that defines a user browsing session to extract data from the Web. Agent Toolbox allows a user to produce an agent simply by browsing the target Web site using the browser embedded in the user interface of Agent Toolbox to provide training examples of a user session, and Agent Toolbox will generalize the training examples into a script. Agent Toolbox virtually converts a Web server into a structured database, facilitating integration of Web data sources. With Agent Toolbox, I developed 5 integrated Web-based services of biomedical data. The most remarkable one is a Web-based SNP prioritization tool called “FastSNP” [c]. FastSNP prioritizes SNPs according to twelve phenotypic risks and putative functional effects. A unique feature of FastSNP is that the functional effect information used for SNP prioritization is always up- to-date, because FastSNP extracts the information from 11 external web servers, such as dbSNP, GeneBank, HapMap, etc., at query time by a team of Web wrapper agents developed and maintained using Agent Toolbox. Users have published more than 500 articles about their association studies using FastSNP. One of the most remarkable outcomes is a Warfarin dosage prediction kit that was later approved for clinical use in 2006. This kit is based on biomarkers identified from a cohort in Taiwan using the software tool FastSNP that my team developed. FastSNP was included in a textbook and in the Wikipedia entry "SNP annotation." Other remarkable Web-based services include EHCO, an integrated database of differentially expressed genes in liver cancer cells [d], and GNS, a database integrating 11 different human gene nomenclatures.

a. Hsu CN, Dung M. Generating finite-state transducers for semi-structured data extraction from the web.

Information systems. 1998; 23(8):521-538. (625 citations, Google Scholar, June 14, 2019)

b. Chang C, Hsu C, Liu S. Automatic information extraction from semi-structured web pages by pattern discovery. Decision Support Systems. 2003 April; 35(1):129-147.

c. Yuan HY, Chiou JJ, Tseng WH, Liu CH, Liu CK, Lin YJ, Wang HH, Yao A, Chen YT, Hsu CN. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W635-41. PubMed PMID: [16845089](http://www.ncbi.nlm.nih.gov/pubmed/16845089/); PubMed Central PMCID: [PMC1538865](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1538865/). (545 citations, Google Scholar, June 14, 2019)

d. Hsu CN, Lai JM, Liu CH, Tseng HH, Lin CY, Lin KT, Yeh HH, Sung TY, Hsu WL, Su LJ, Lee SA, Chen CH, Lee GC, Lee DT, Shiue YL, Yeh CW, Chang CH, Kao CY, Huang CY. Detection of the inferred interaction network in hepatocellular carcinoma from EHCO (Encyclopedia of Hepatocellular Carcinoma genes Online). BMC Bioinformatics. 2007 Feb 27;8:66. PubMed PMID: [17326819](http://www.ncbi.nlm.nih.gov/pubmed/17326819/); PubMed Central PMCID: [PMC1828168.](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1828168/)

**4. Cellular Image Analysis for Drug Screening**

As a project leader in the Advanced Bioinformatics Core of the National Research Program on Genomic Medicine in Taiwan, one of my contributions is cellular image analysis for high-throughput screening, including a tool that enabled us to reveal new roles of two proteins regulating mitochondrial dynamics [c].

a. Lin CC, Tsai YS, Lin YS, Chiu TY, Hsiung CC, Lee MI, Simpson JC, Hsu CN. Boosting multiclass learning with repeating codes and weak detectors for protein subcellular localization. Bioinformatics. (ISMB) 2007 Dec 15;23(24):3374-81. PubMed PMID: 17956879.

b. Lin YS, Lin CC, Tsai YS, Ku TC, Huang YH, Hsu CN. A spectral graph theoretic approach to quantification and calibration of collective morphological differences in cell images. Bioinformatics (ISMB). 2010 Jun 15;26(12):i29-37. PubMed PMID: 20529919; PubMed Central PMCID: PMC2881379.

c. Peng JY, Lin CC, Chen YJ, Kao LS, Liu YC, Chou CC, Huang YH, Chang FR, Wu YC, Tsai YS, Hsu CN. Automatic morphological subtyping reveals new roles of caspases in mitochondrial dynamics. PLoS Comput Biol. 2011 Oct;7(10): e1002212. PubMed PMID: 21998575; PubMed Central PMCID: PMC3188504.

**D. Additional Information: Research Support and/or Scholastic Performance**

**Ongoing Research Support**

DM190543 Hsu (PI)

CDMRP/DoD 07/15/20-01/14/22

Deep Learning for Imaging Report Generation to Support Diagnosis of Military-Relevant Injury in a Deployed or Operational Environment

The goal of the project is to develop a deep neural network system to generate a radiology report from an orthopedic radiograph of musculoskeletal injury in the battlefield.

Role: PI

2 U24DK097771-06 Grethe, Martone (MPI) 07/01/18 – 06/30/23

NIDDK/NIH

DKNET Coordinating Unit: An Information Network for FAIR Resources and Data

This project aims at the establishment of the NIDDK Interconnectivity Network Coordinating Center (INCC) to expand and enhance the current NIDDK Consortium Interconnectivity Network (dkCOIN) community and infrastructure, providing seamless access to large pools of data relevant to the mission of NIDDK.

Role: Project Scientist

Multi-Year Project 01/01/18 – 12/31/22

IBM Corp.

IBM Cognitive Horizons Network on AI for Healthy Living

A multi-year project in two thematic areas: Healthy Aging and the Human Microbiome. This collaboration is part of the [IBM Cognitive Horizons Network](http://research.ibm.com/cognitive-computing/cognitive-horizons-network/), an international [consortium of 9 universities](https://www.ibm.com/blogs/research/2017/09/advancing-ai/) working with IBM to develop technologies needed to help fulfill the promise of artificial intelligence (AI).

Role: Nature Language Processing Group Leader

**Completed Research Support**

Chancellor’s Interdisciplinary Collaboratories Fellowships Hsu, McAuley, et al. (MPI) 01/01/18 – 12/31/19

UC San Diego

Training conversational agents with clinical text for Neurology patient intake and evaluation. The goal is to develop deep learning algorithms to train conversational agents for neurology patient intake and evaluation using the text data from clinical notes and documents stored in an electronic medical record system.

Role: PI

CDRN-1306-04819 Ohno-Machado (PI) 03/31/14 - 09/30/18

PCORI

pSCANNER: Patient-Centered SCAlable National Network for Effectiveness Research

This project will allow distributed queries across the VA national enterprise data warehouse, the five University of California medical center clinical data warehouses, and three federally qualified health systems in LA.

Role: Co-Investigator, Workgroup Chair, Natural Language Processing (NLP).

5 U01 HG006894-04 Hsu, Chun-Nan (PI) 09/24/12-09/30/16

NIH/NHGRI

Accelerating Curation of GWAS Catalog by Automatic Text Mining

The major goals of this project are to perform automatic text mining of current genetics literature to establish a highly curated catalog of published genome-wide association studies.

Role: PI

4 UH3 HL108785-03 Ohno-Machado, Lucila (PI) 09/24/13-09/23/15

NIH/NHLBI

Phenotype Discovery in NHLBI Genomic Studies (PhD)

The major goal of this project is to build an automated tool for standardizing phenotypic and genotypic

annotation of data from genomic public repositories, such as dbGaP.

Role: Co-Investigator

**Peer Reviewed Publications**

Complete lists of publications can be found at:

My Bibliography at NCBI: <https://www.ncbi.nlm.nih.gov/myncbi/chun-nan.hsu.1/bibliography/public/>