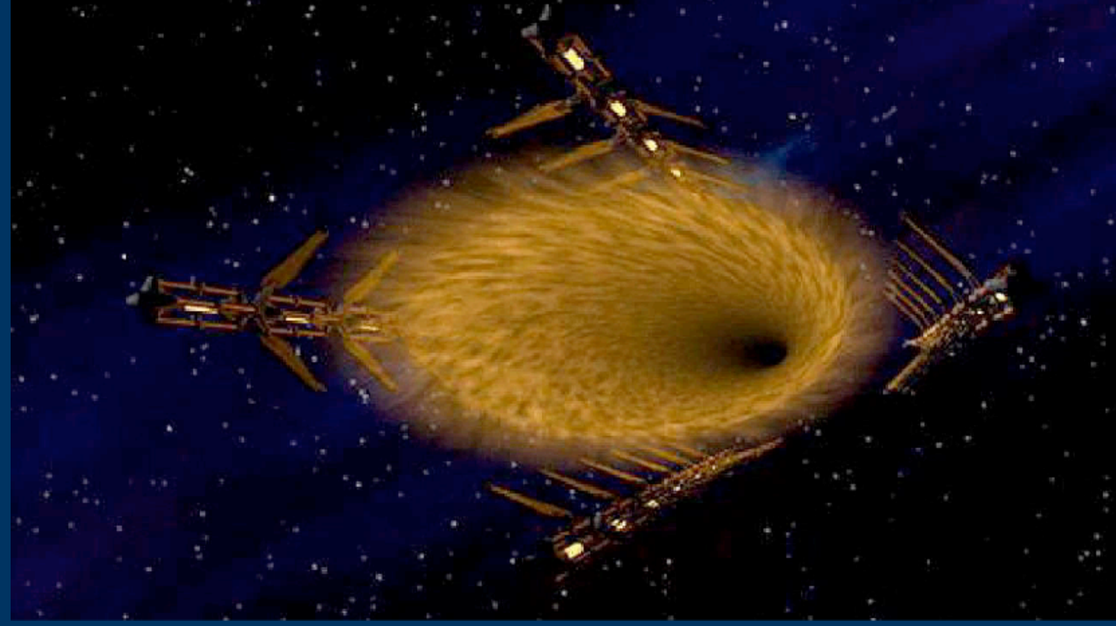
my bioinformatics journey

career talk for Erdős Institute



Andrew Uzilov, PhD
Senior Director (of) Computational Biology (in) the Discovery Research group (at) Veracyte, Inc.
2025-09-29
https://www.linkedin.com/in/andrewuzilov/

Outline

- What is “bioinformatics” ?
- My career journey
- What we do at Veracyte
- Advice

Disclaimer

These opinions are mine.

Therefore, they are biased towards my personal experience as a:

- bioinformatics scientist specializing in sequence analysis and NGS
- mainly in cancer genomics
- in the diagnostics industry
- who often collaborates with people in other disciplines

What is “bioinformatics”, anyway?

What is bioinformatics like?

- There are no good catch-all answers because there are many types of bioinformatics
- Inherently inter-disciplinary but also hacky
- Bioinformatics started in ***sequence analysis***
  - evolution of the field is influenced by the how much sequencing data is produced and on computational power
  - earlier work is foundational algorithms and small-dataset examples of utility

J. Mol. Biol. (1970) 48, 443–453

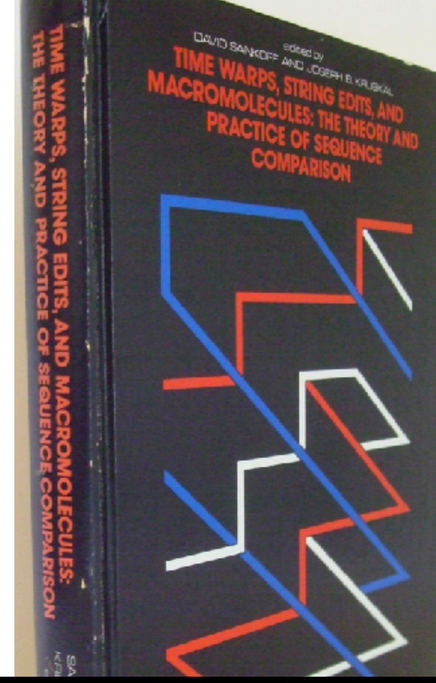
A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

SAUL B. NEEDLEMAN AND CHRISTIAN D. WUNSCH

Department of Biochemistry, Northwestern University, and
Nuclear Medicine Service, V. A. Research Hospital
Chicago, Ill. 60611, U.S.A.

(*Received 21 July 1969*)

A computer adaptable method for finding similarities in the amino acid sequences of two proteins has been developed. From these findings it is possible to determine whether significant homology exists between the proteins. This information is used to trace their possible evolutionary development.

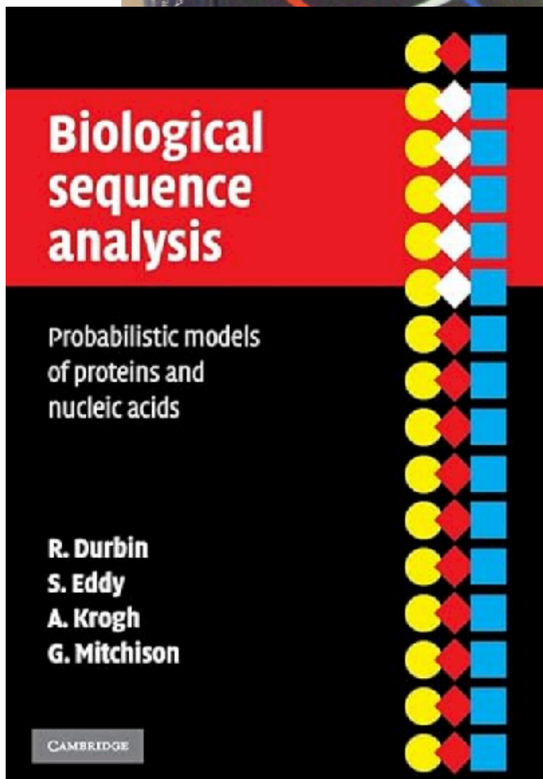


SIAM J. APPL. MATH.
Vol. 45, No. 5, October 1985

SIMULTANEOUS SOLUTION OF THE RNA FOLDING, ALIGNMENT AND PROTOSEQUENCE PROBLEMS*

DAVID SANKOFF†

Abstract. The alignment of finite sequences, the inference of ribonucleic acid secondary structures (folding), and the reconstruction of ancestral sequences on a phylogenetic tree, are three problems which have dynamic programming solutions, which we formulate in a common mathematical framework. Combining the objective functions for alignment (parsimony, or minimal mutations) and folding (free energy), we present an algorithm which solves all three problems simultaneously for a set of N sequences of length n in time proportional to n^{3N} and storage n^{2N}. Incorporating a "cutting corners" constraint against biologically unlikely alignments reduces these requirements so that they are proportional to n^3 and n^2, respectively, for fixed N.



What are the different types of bioinformatics?

As with any scientific discipline, it depends on the *methods* you use

- quantitative transcriptomics (NGS or expression microarrays)
- algorithms
- network modeling (gene/regulatory networks, protein/protein interaction networks)
- protein or RNA structure (NMR/X-ray crystallography, computational structure modeling)
- Real World Data (RWD): ETL, NLP, EMR/EHR, LLM/AI, sparse+messy+huge databases
- knowledge modeling (ontologies, knowledge graphs)

sequence analysis: 

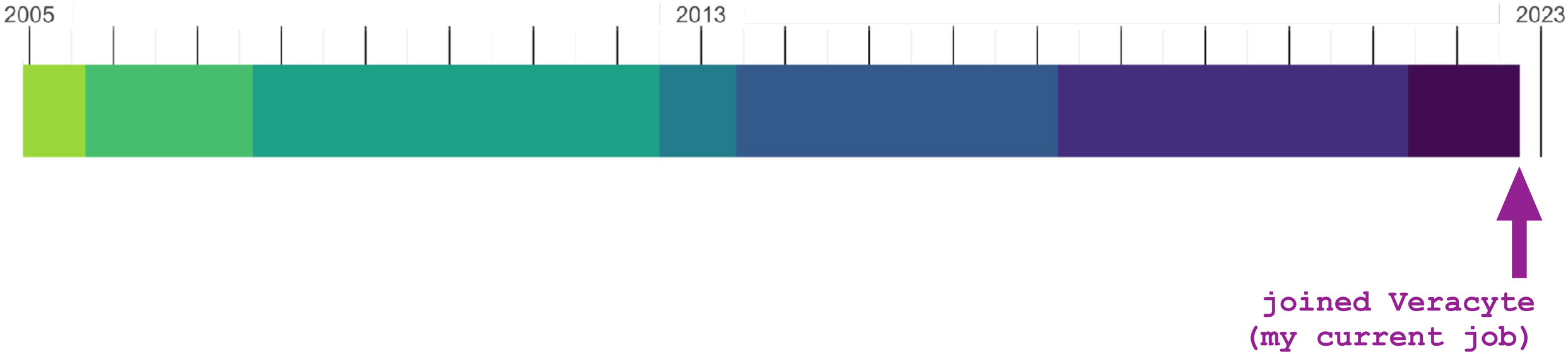
- secondary analysis
  - basically, from FASTQ to everything else
  - alignment, QC, variant calling, complex genomic biomarkers (CNV, TMB, HRD, MSI, kataegis, mutational signatures), annotation
- genome assembly
- molecular phylogeny

Trends over the past 2 decades

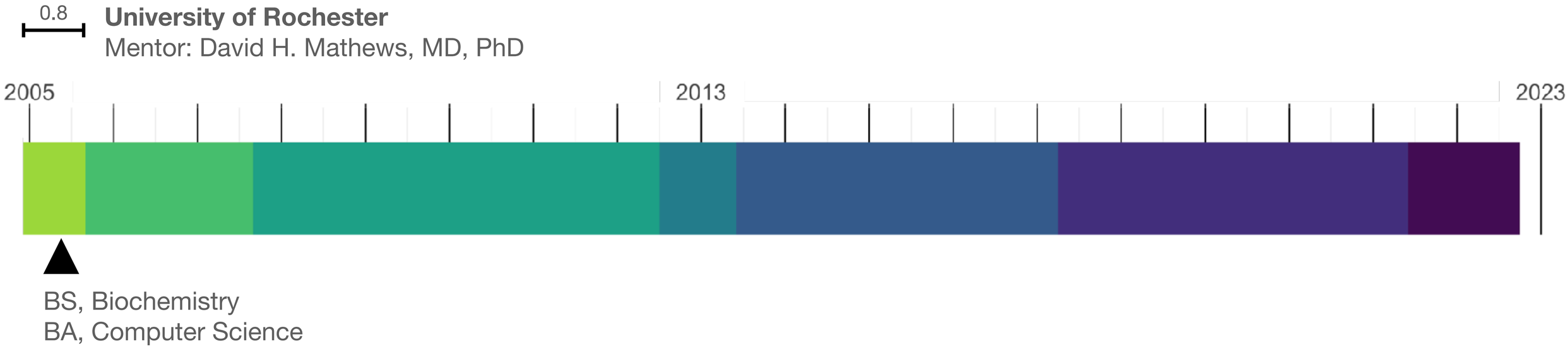
- Bioinformatics is increasingly a skillset combined with other specialties, not a career itself. Problem-solving is becoming more commoditized. It pays to invest in combining with other fields.
- Tool standardization: R/Bioconductor, the Broad ecosystem, UCSC Genome Browser
- Emergence of viable commercial solutions shows the industry is reaching maturity: DNANexus, Velsera, Sophia Genetics, Sentieon, AWS HealthOmics, Dragen, Illumina's crapware (BaseSpace, ICA, ICI), etc…
- NGS instrumentation as an industry: Illumina, Ion Torrent (Thermo Fisher), PacBio, Oxford Nanopore, Complete Genomics, and now in the 2020s… Ultima, Element, Singular, Roche SBX
- One thing held true through my entire career… most people **do not understand what it is a "bioinformatician" actually does!**

My career journey, step by step

Career Timeline



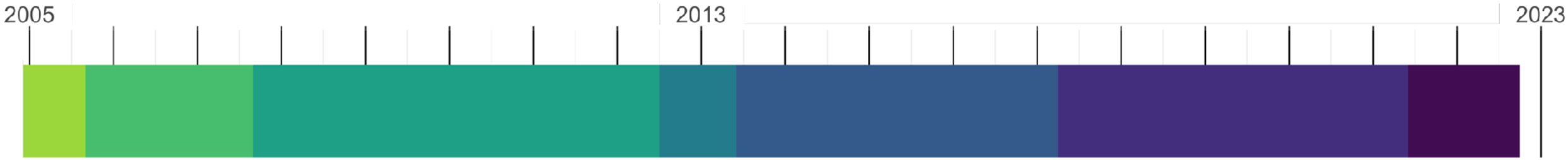
Career Timeline

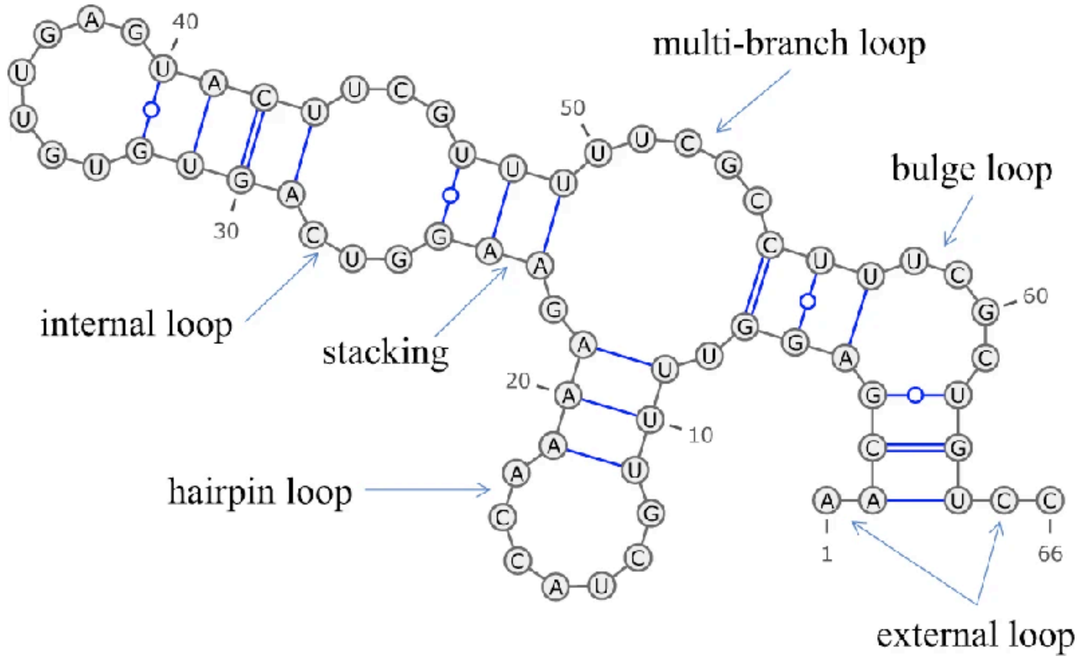


Career Timeline

0.8

University of Rochester
Mentor: David H. Mathews, MD, PhD





Research article | Open Access | Published: 27 March 2006

Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change

Andrew V Uzilov, Joshua M Keegan & David H Mathews

BMC Bioinformatics **7**, Article number: 173 (2006) | Cite this article

11k Accesses | **131** Citations | **1** Altmetric | Metrics

undergraduate research →
my first peer-reviewed paper!

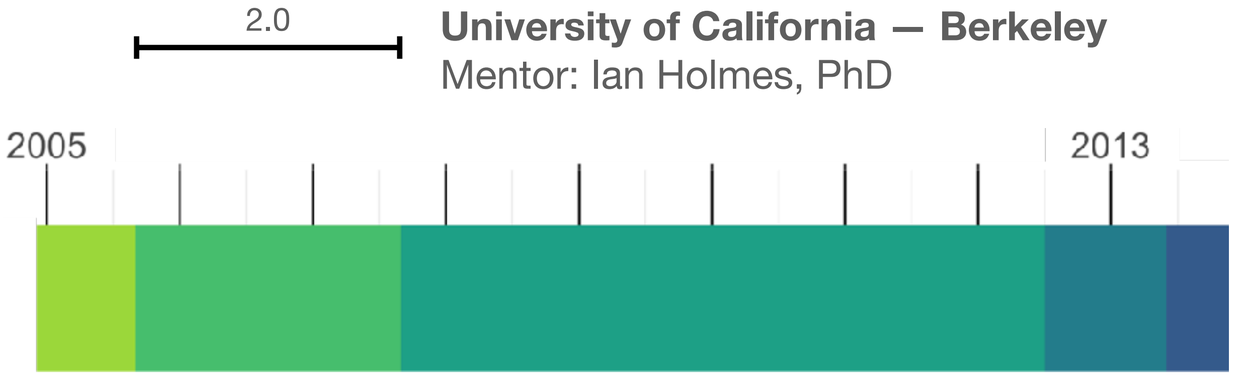
https://doi.org/10.1038/s41467-021-21194-4

annual salary: ~$30K

https://doi.org/10.1186/1471-2105-2-8

https://doi.org/10.1186/1471-2105-5-71

Career Timeline



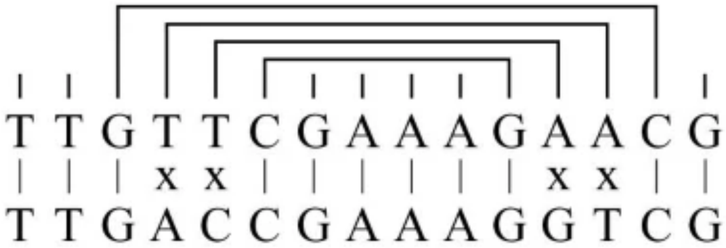
University of California — Berkeley
Mentor: Ian Holmes, PhD

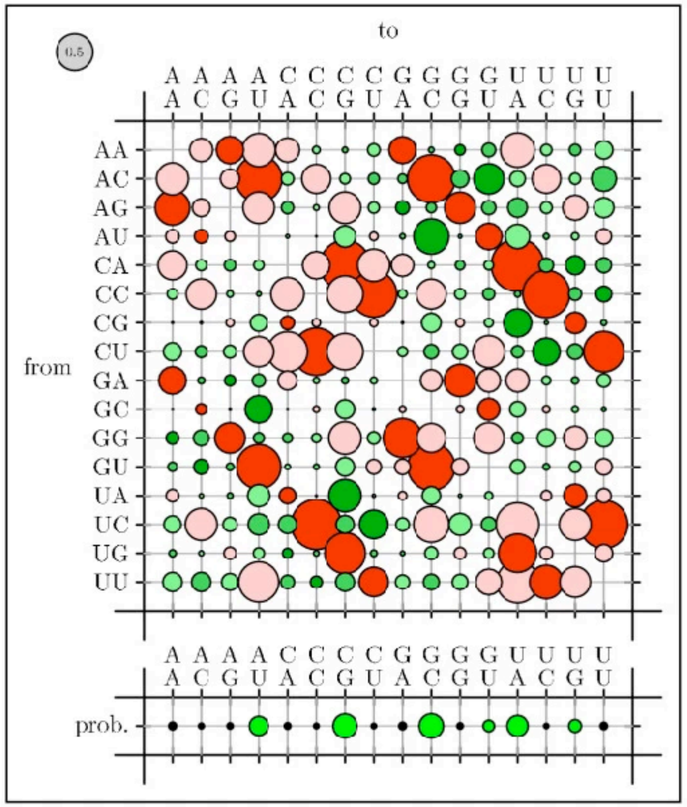
G3: $S \to aS\hat{a} \mid aL \mid Ra \mid LS$
$L \to aS\hat{a} \mid aL$
$R \to Ra \mid \varepsilon$

G4 : $S \to aS \mid T \mid \varepsilon$
$T \to Ta \mid aS\hat{a} \mid TaS\hat{a}$

G5 : $S \to aS \mid aS\hat{a}S \mid \varepsilon$

G6 : $S \to LS \mid L$
$L \to aF\hat{a} \mid a$
$F \to aF\hat{a} \mid LS$





Software | Open Access | Published: 03 October 2006

XRate: a fast prototyping, training and annotation tool for phylo-grammars

Peter S Klosterman, Andrew V Uzilov, Yuri R Bendaña, Robert K Bradley, Sharon Chao, Carolin Kosiol, Nick Goldman & Ian Holmes

BMC Bioinformatics **7**, Article number: 428 (2006) | Cite this article

10k Accesses | **43** Citations | **3** Altmetric | Metrics

annual salary: ~$30K

https://jbrowse.org

Career Timeline

2.0

University of California — Berkeley
Mentor: Ian Holmes, PhD

2005 — 2013

CSH PRESS GENOME RESEARCH

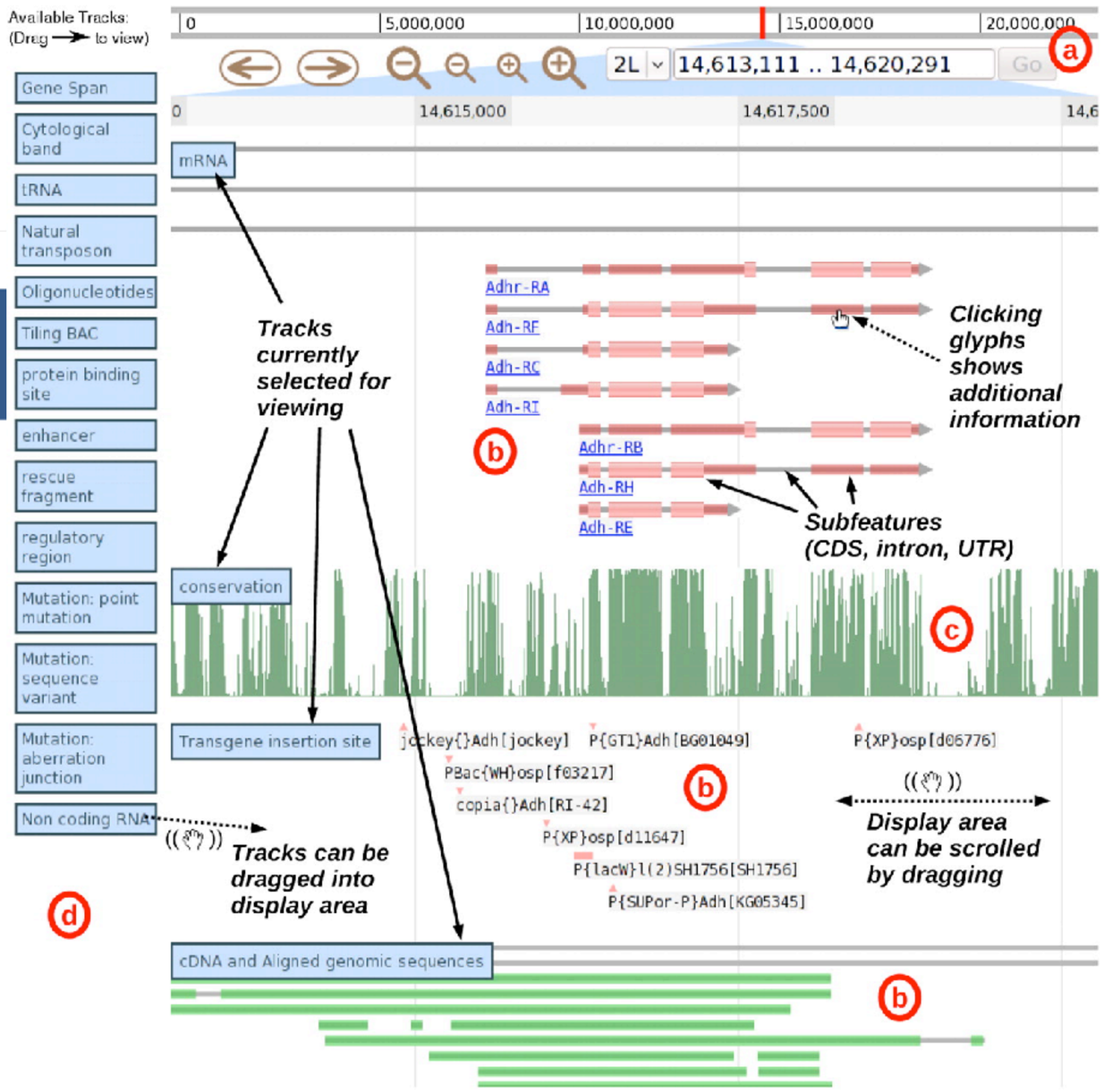
JBrowse: A next-generation genome browser

Mitchell E. Skinner[1], Andrew V. Uzilov[1], Lincoln D. Stein[2], Christopher J. Mungall[3] and Ian H. Holmes[1,3,4]

OPEN ACCESS ARTICLE

This Article

Published in Advance July 1, 2009, doi: 10.1101/gr.094607.109

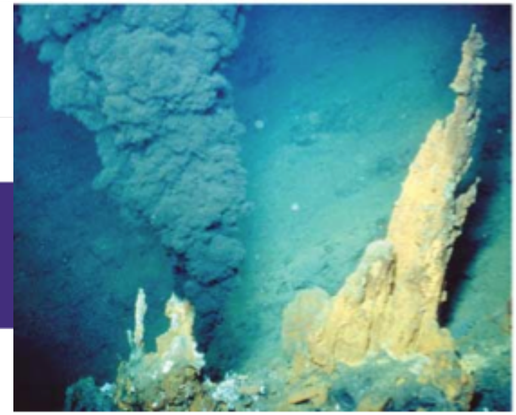


annual stipend: ~$30K under the NSF GRFP fellowship

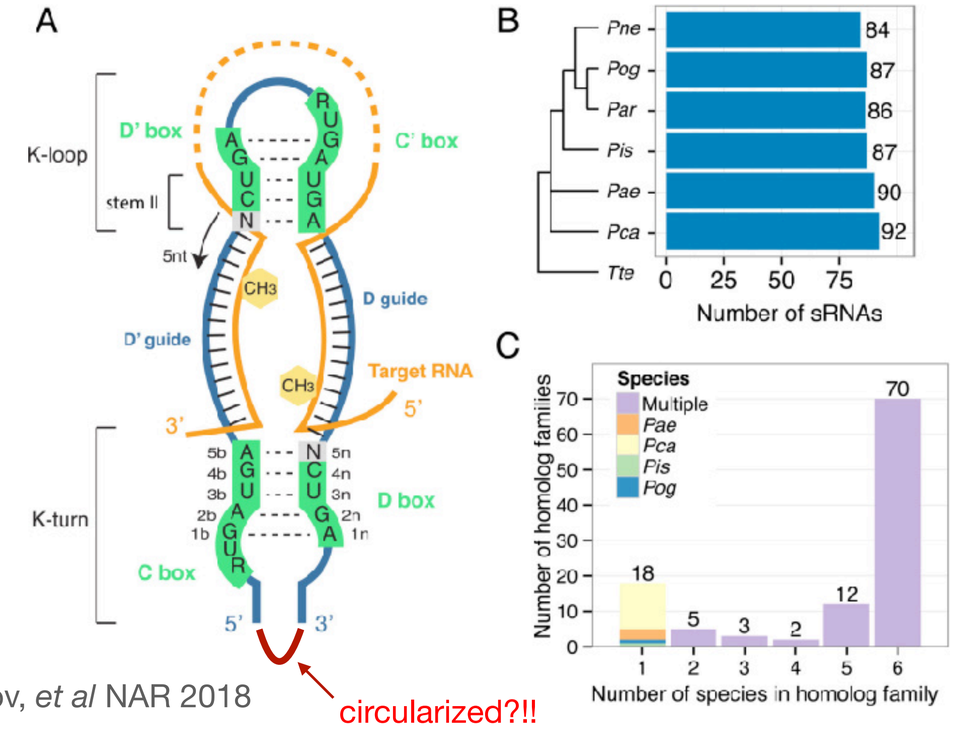
Career Timeline

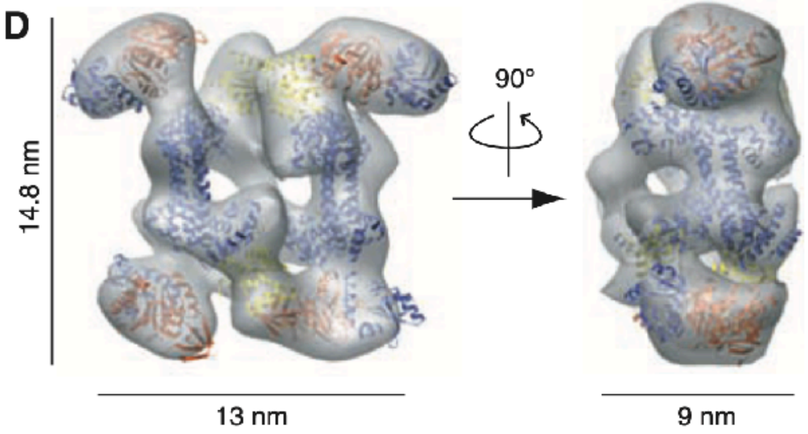
4.8

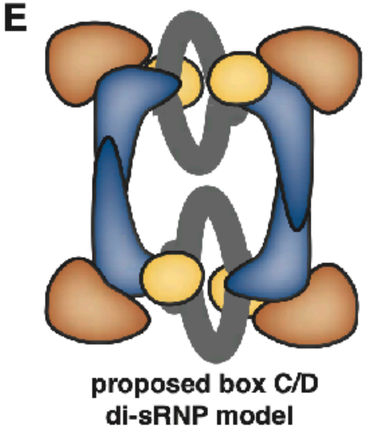
University of California — Santa Cruz
Mentor: Todd M. Lowe, PhD



2005 — 2013





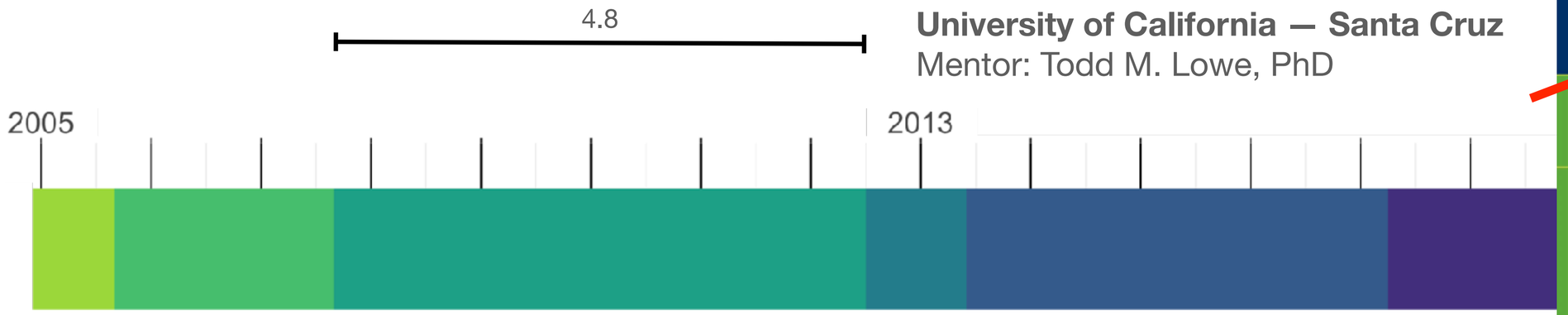


Lui, Uzilov, *et al* NAR 2018

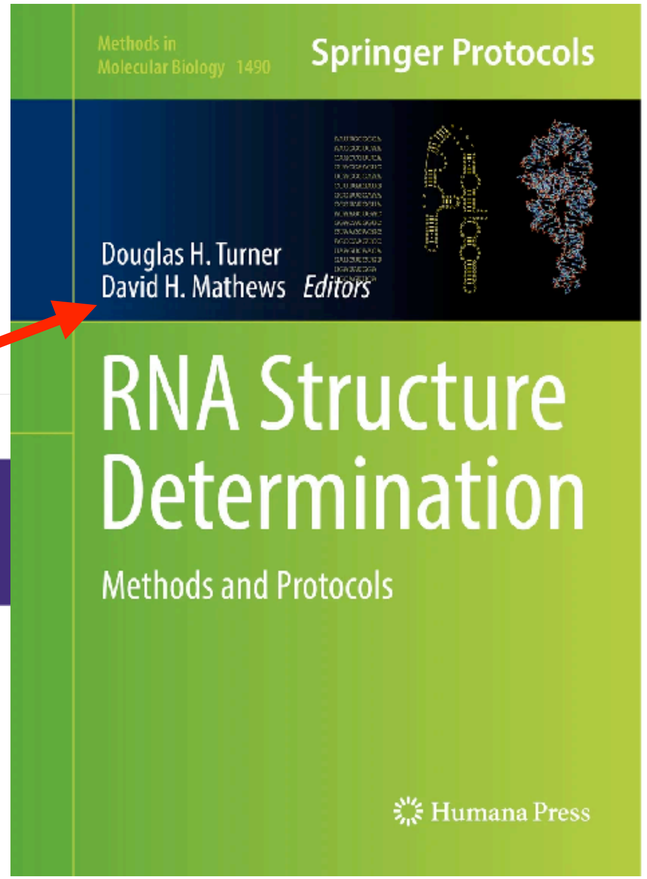
https://doi.org/10.1126/science.1176099

annual stipend: ~$30K under the NSF GRFP fellowship

Career Timeline



University of California — Santa Cruz
Mentor: Todd M. Lowe, PhD



FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing

Jason G Underwood, Andrew V Uzilov, Sol Katzman, Courtney S Onodera, Jacob E Mainzer, David H Mathews, Todd M Lowe, Sofie R Salama & David Haussler

Nature Methods **7**, 995–1001 (2010) | Cite this article

4702 Accesses | **238** Citations | **14** Altmetric | Metrics

networking pays off!

The RNA structurome: high-throughput probing

Eric Westhof & Pascale Romby

Nature Methods **7**, 965–967 (2010) | Cite this article

1150 Accesses | **25** Citations | Metrics

I forgot how much they paid me, but it was part-time, in addition to my graduate stipend

Career Timeline

0.9

transition period:
UCSC, Cancer Therapeutics Innovation Group (aka Driver), etc.

2005 — 2013 — 2023

PhD, Bioinformatics and
Biomolecular Engineering

Journal of Clinical Oncology > List of Issues > Volume 31, Issue 15_suppl >

Meeting Abstract | 2013 ASCO Annual Meeting I

TUMOR BIOLOGY

Integrated genomic analysis by whole exome and transcriptome sequencing of tumor samples from EGFR-mutant non-small-cell lung cancer (NSCLC) patients (p) with acquired resistance to erlotinib.

Jonathan S. Weissman, Petros Giannikopoulos, John St. John, Andrew V. Uzilov, Carlota Costa, Niki KarachaliouIrene Sansano, Eloisa Jantus-Lewintre, Rolf A. Stahel, Alain Vergnenegre, Radj Gervais, Jose Luis Perez-Gracia, Maria D. Lozano, Anne S. Wellde, Rodolfo Bordoni, Andres Felipe Cardona Zorrilla, William Reilly Polkinghorn, George W. Wellde, Rafael Rosell, Trever Grant Bivona.

ARTICLE CITATION

DOI: 10.1200/jco.2013.31.15_suppl.11010
Journal of Clinical Oncology 31, no. 15_suppl
(May 20, 2013) 11010-11010.

Journal of Clinical Oncology > List of Issues > Volume 31, Issue 15_suppl >

Meeting Abstract | 2013 ASCO Annual Meeting I

TUMOR BIOLOGY

Integrated genomic analysis of EGFR-mutant non-small cell lung cancer immediately following erlotinib initiation in patients.

Trever Grant Bivona, Petros Giannikopoulos, Carlota Costa, Niki Karachaliou, Santiago Viteri, M. Rosario Garcia-CampeloJohn St. John, Andrew V. Uzilov, Anne S. Wellde, William Reilly Polkinghorn, Margarita Majem, Enriqueta Felip, Enric Carcereny, Cordula Nicole Heidecke, Bartomeu Massuti, George W. Wellde, Jonathan S. Weissman, Rafael Rosell.

ARTICLE CITATION

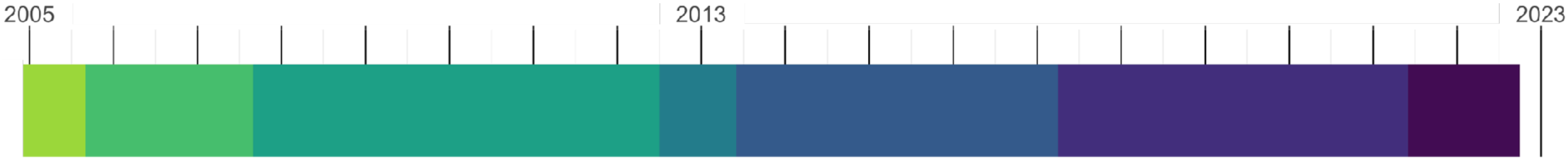
DOI: 10.1200/jco.2013.31.15_suppl.11067
Journal of Clinical Oncology 31, no. 15_suppl
(May 20, 2013) 11067-11067.

base salary at hiring: $135K/year, no annual bonus

Career Timeline

3.8

Icahn School of Medicine at Mount Sinai
Lab of Rong Chen, PhD



Research | Open Access | Published: 01 June 2016

Development and clinical application of an integrative genomic approach to personalized cancer therapy

Andrew V. Uzilov, Wei Ding, Marc Y. Fink, Yevgeniy Antipin, Andrew S. Brohl, Claire Davis, Chun Yee Lau, Chetanya Pandya, Hardik Shah, Yumi Kasai, James Powell, Mark Micchelli, Rafael Castellanos, Zhongyang Zhang, Michael Linderman, Yayoi Kinoshita, Micol Zweig, Katie Raustad, Kakit Cheung, Diane Castillo, Melissa Wooten, Imane Bourzgui, Leah C. Newman, Gintaras Deikus, Bino Mathew, Jun Zhu, Benjamin S. Glicksberg, Aye S. Moe, Jun Liao, Lisa Edelmann, Joel T. Dudley, Robert G. Maki, Andrew Kasarskis, Randall F. Holcombe, Milind Mahajan, Ke Hao, Boris Reva, Janina Longtine, Daniela Starcevic, Robert Sebra, Michael J. Donovan, Shuyu Li, Eric E. Schadt ✉ & Rong Chen ✉

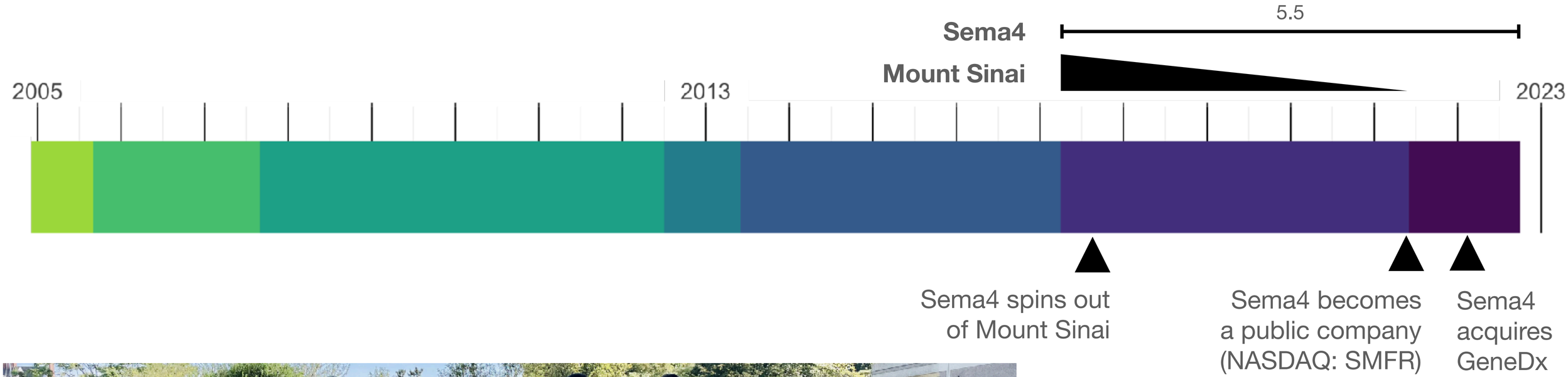
Genome Medicine **8**, Article number: 62 (2016) | Cite this article

10k Accesses | **56** Citations | **551** Altmetric | Metrics

Mount Sinai

Icahn Institute for Genomics and Multiscale Biology

Career Timeline



Somatic Genomics R&D (SGRD) group at its peak in 2021

Career Timeline

ELSEVIER

European Urology

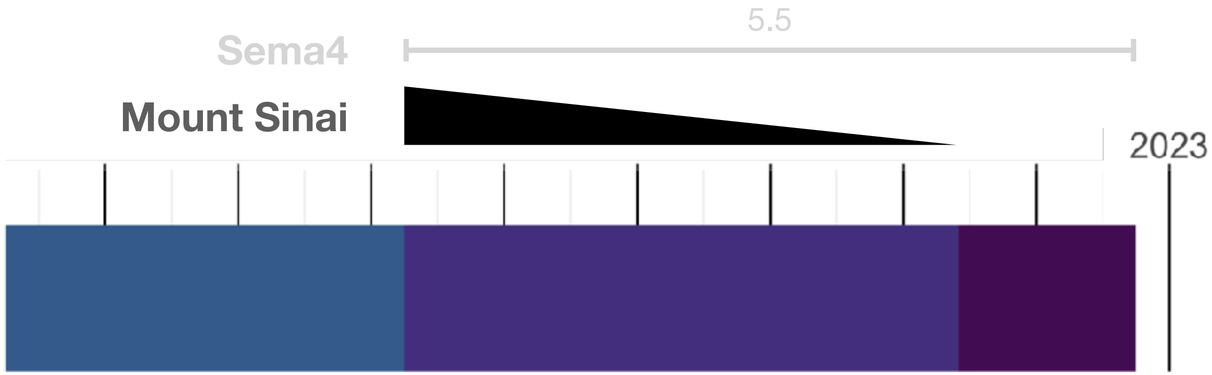
Volume 73, Issue 5, May 2018, Pages 751-759

Platinum Priority – Urothelial Cancer

Editorial by Daniel M. Geynisman, Phillip H. Abbosh, Elizabeth R. Plimack and Matthew Zibelman on pp. 760–762 of this issue

Phase 2 Trial of Gemcitabine, Cisplatin, plus Ipilimumab in Patients with Metastatic Urothelial Cancer and Impact of DNA Damage Response Gene Mutations on Outcomes

Matthew D. Galsky [a], Huan Wang [b, c], Noah M. Hahn [d], Przemyslaw Twardowski [e], Sumanta K. Pal [c], Costantine Albany [f], Mark T. Fleming [g], Alexander Starodub [h], Ralph J. Hauke [i], Menggang Yu [j], Qianqian Zhao [j], Guru Sonpavde [k], Michael J. Donovan [l], Vaibhav G. Patel [a], John P. Sfakianos [m], Josep Domingo-Domenech [l], William K. Oh [a], Nicholas Akers [b], Bojan Losic [b], Sacha Gnjatic [a], Eric E. Schadt [b, c], Rong Chen [b, c], Seunghee Kim-Schulze [a], Nina Bhardwaj [a], Andrew V. Uzilov [b, c]



RESEARCH HIGHLIGHTS

Nature Reviews Urology | Published online 3 Jan 2017; doi:10.1038/nrurol.2017.224

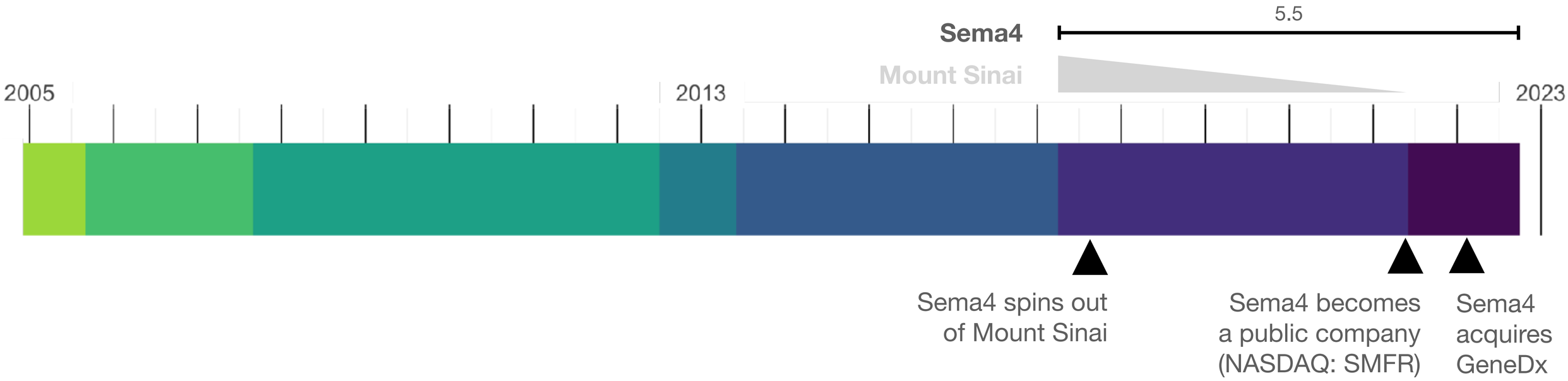
BLADDER CANCER

Chemotherapy and checkpoint blockade

A phase II trial of gemcitabine and cisplatin (GC) plus ipilimumab in 36 chemotherapy-naive patients with metastatic urothelial cancer did not 'autovaccination' effect caused by chemotherapy, which might combine favourably with CTLA4 blockade." Using a phased treatment schedule (two cycles of GC followed by four cycles of GC plus ipilimumab), the team evaluated whether immunosuppressive effects of chemotherapy might hamper its immunomodulatory effects, and whether responses depend on a specific genomic tumour make up. "We hypothesized that tumours with DDR not limit the activity of immune checkpoint blockade. Deleterious tumour DDR gene mutations were associated with increased total mutational burden and a significantly increased likelihood of response ($P = 0.03$); ten patients with these alterations had a partial or complete response. Testing whether GC induces immunogenic cell death, the researchers found no increase in levels of HMGB1, a danger-associated signal

Career Timeline

base annual salary (basically funded by Sema4):
~$170K as Director
~$200K as Senior Director
~$260K as VP (+ ~$40K annual bonus?)
plus ~$0.5M (at peak) in stock options that I couldn't cash out, earning $0
plus some RSUs after we went public



Sema4 Signal Hereditary Cancer (HC)

Sema4 Signal Whole Exome/Transcriptome Sequencing for all cancers (WES/WTS)

Sema4 Signal Solid Tumor Panel

Clinical tests

to which I majorly contributed at Sema4

All tests are NYSDOH approved and have been in clinical operation for 2-4 years, resulting >10K patients combined. Also used to support major pharma projects.

Sema4 Signal Hereditary Cancer (HC)

- Normal DNA (for germline/ constitutional variants)
- 113 HC genes on a proprietary “medical exome” hyb capture panel
- SNV and small ins/del
- CNV to exon resolution
- QC

Sema4 Signal Whole Exome/Transcriptome Sequencing for all cancers (WES/WTS)

- Tumor/normal DNA
- Tumor-only RNA
- Solid and heme tumors
- ~18.5K genes for WES
- RNA-Seq workflows: hyb capture of cDNA vs total RNA
- SNV, small ins/del (LOD AF 5%)
- Genome-wide CNV by segmentation
- MSI and TMB
- Fusions and certain splice variants from RNA (expression/ abundance RUO)
- QC

Sema4 Signal Solid Tumor Panel

- Tumor-only
- Solid tumors (no heme)
- Ion AmpliSeq targeted panel (amplicons, multiplex PCR)
- 161 genes (141 DNA, 25 RNA, ~0.3 Mb)
- SNV, small ins/del/MNV/delins
- CNV (amplifications only)
- Fusions and certain splice variants from RNA
- QC
- Annotation and variant prioritization (tertiary analysis)

My role:
Ownership of all NGS secondary analysis
- assay development and optimization
- pipelines: R&D and CLIA

What we do at Veracyte

What does Veracyte do?

- We offer clinical diagnostic tests that use machine learning classifiers on patient **RNA** to give **diagnosis and risk stratification to the physician**
  - RNA-seq, expression microarrays, NanoString nCounter, qPCR
- We save tens of thousands of patients from having their organs unnecessarily removed! (ex: ~35k thyroids per year)
- Launching a **DNA test**: an all-WGS MRD test (via acquiring C2i Genomics)
- We also do a great job of:
  - demonstrating our products are useful via clinical studies
  - collaborating with academics using our gigantic NGS datasets

Afirma
THYROID FNA ANALYSIS

VV20008

REPORT STATUS: Final
PAGES: 1 of 2
CLIENT ID: 97
AFIRMA REQ: R123

PATIENT REPORT

Sample Patient Report

PATIENT INFORMATION

PATIENT: John Doe | **DOB:** 01 Jan 1973 | **GENDER:** M | **LAB ID:** L123 | **MRN:** M123

COLLECTION DATE	18 Sep 2019	**FACILITY NAME**	University Hospital of Anytown		
RECEIVED DATE	20 Sep 2019	**SUBMITTING PHYSICIAN**	Jane Demo	**PHONE**	(555) 555-5555
REPORT DATE	26 Sep 2019	**TREATING PHYSICIAN/CC**	---	**PHONE**	---

CLINICAL HISTORY: Suspicious Ultrasound Characteristics: Nodule A: Hypoechoic, Solid: >95% solid

RESULTS

Nodule: A 2.1 cm, Middle Left

CYTOPATHOLOGY

I	II	III	IV	V	VI
Non Diagnostic	Benign	Atypia of Undetermined Significance	**Suspicious for Follicular Neoplasm**	Suspicious for Malignancy	Malignant

Cytopathology Diagnosis: Indeterminate - Suspicious for Follicular Neoplasm

Diagnostic Comments: These features are best categorized as suspicious for follicular neoplasm, Hürthle cell type.

Microscopic Description: The cytologic preparations are highly cellular and predominantly contain Hurthle cells in single cells or crowded groups. Several of the Hurthle cells are enlarged and show round, pale nuclei. Some colloid and relatively few lymphocytes are also seen.

AFIRMA GENOMIC SEQUENCING CLASSIFIER

Benign
(Risk of Malignancy ~4%)

MTC: Negative
Parathyroid: Negative

***BRAF*:p:V600E c. 1799T>A:** Negative
***RET/PTC1, RET/PTC3*:** Not Detected

AFIRMA XPRESSION ATLAS

N/A

RESULTS INTERPRETATION

The result of this 2.1 cm Bethesda IV nodule A is Afirma GSC Benign, which suggests a low risk of cancer at approximately 4%. Treatment like a cytologically benign nodule may be appropriate, including clinical correlation. Afirma XA is not performed on GSC Benign nodules.[7]

base annual salary as Senior Director: $270-290K
+ ~$60K annual bonus (25% bonus target)
+ ~$160K in new RSUs per year (1-year vest, then disbursed over 4 years)
(bonus and RSUs vary by year depending on company performance

What do I do?

title: Senior Director, Computational Biology, Discovery Research

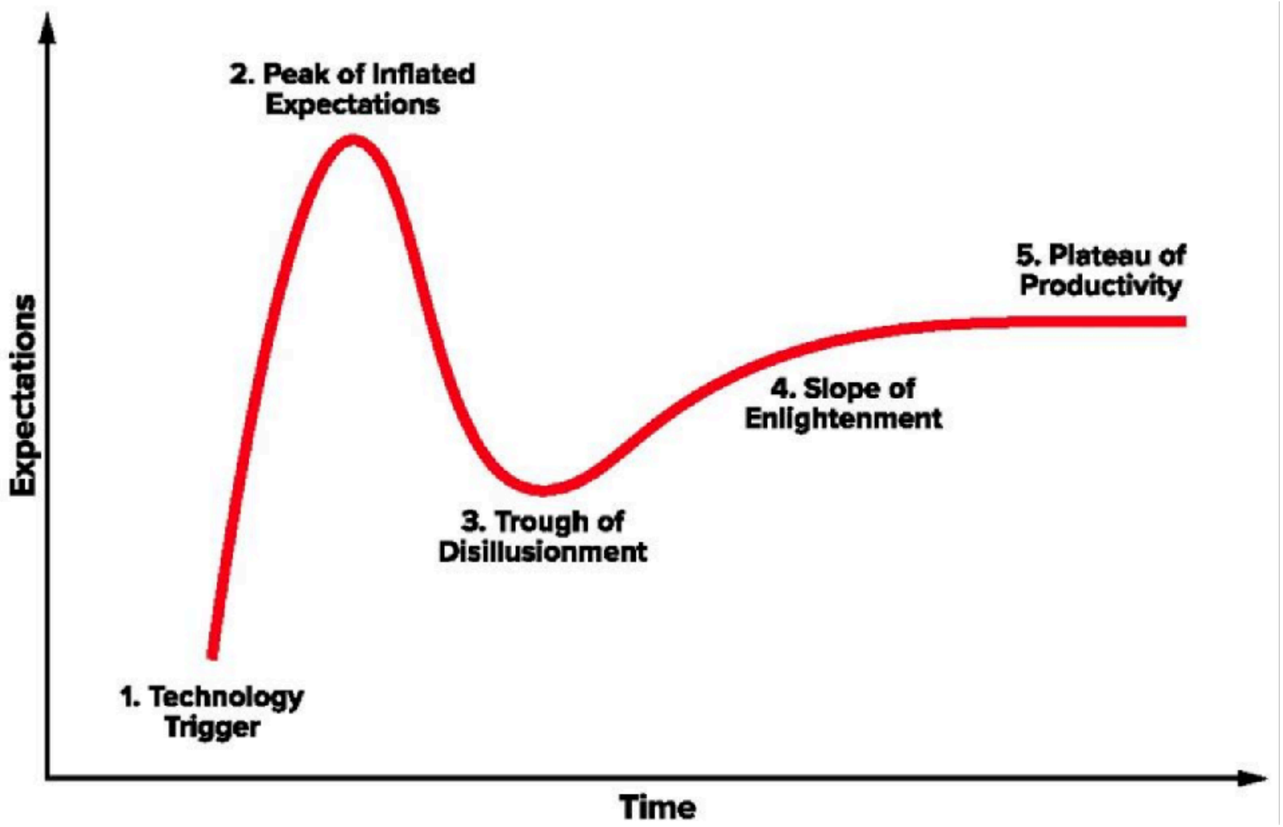
- I make educated guesses about what we should develop next as a clinical diagnostics company and build up a case for it
- Basically, I “discover” new stuff and show that it is useful to the company
- We do exploratory, feasibility, proof-of-concept, and those types of studies
  - internally by standing up new assays in our lab
  - externally by doing pilots with stealth mode startups and academic parties
  - also strictly dry lab studies re-analyzing our treasure troves of NGS data and RWD

Advice to gradute students and postdocs

There will always be hype

Hype waves that I've been through

- "assembling the human genome will solve everything"
- "precision medicine will solve everything"
- "data science will solve everything"
- "generative AI will solve everything"
- "multimodal analysis", a next phase beyond "multiomics", will "unveil new insights" **<— WE ARE HERE**



Opportunities to test the waters in an industry career

- Internships (!!!)
- Mentorship
- Shadowing
- Random conversations you start with people on LinkedIn
- Industry postdocs

What you can do to develop bioinformatics skills

- The Holy Trinity: R (esp. the “tidyverse” library) // Python (v3.x) // SQL of any flavor
- AWS (core products: EC2, S3, IAM)
- get a Mac
- computing at scale (“batch computing”) — doing the same thing on a LOT of data in parallel
- generative AI (esp. Claude to write code) can help tremendously
  - however, keep in mind it is essentially search engine (a “glorified parrot”)
  - **fast but bad** is often better to get you *started* than **slow but high quality** as long as you’re a **critical thinker**
- don’t *blindly* trust authority (leadership is incentivized to control the narrative)

Connecting to the right job

things that might not be immediately obvious

- **What I look for when interviewing as a hiring manager**
  - are you a basically binf tool user or is there evidence that you designed and ran your own projects?
  - “amphibious” (both wet and dry lab) background is a HUGE plus
- **What you should look for when interviewing**
  - Do the people interviewing you have a clue about what your job would entail, or the different types of bioinformatics? And that bioinformatics overlaps with, but is different from, other “computational” functions like IT, software development, devops, data engineering?
  - Are you going to be the only bioinformatician there?

Pivoting to science-adjacent careers

for a scientist who decides not to continue doing research

- Field Application Scientist
- Business Development (MBA/PhD)
- Sales/marketing (including Medical Science Liaison)
- IP lawyer (JD/PhD)

Thank You!