XIAOTONG YAO

+1(917)833-5729 ◇ xiy2006@med.cornell.edu ◇ xiaotong.yao23@gmail.com
Github, Twitter: xtYao ◇ ORCID: 0000-0002-5140-6639
1300 York Ave ◇ New York, NY 10065

EDUCATION

Weill Cornell Medicine *Jul 2015- expected Apr 2021*
Ph.D. in Computational Biology & Medicine
Thesis (supervisor: Marcin Imielinski): *Illuminating rearranged cancer genome structures throguh genome graphs*
Relevant courses: Optimization Methods, Data Structure and Algorithms, Poplulation Genetics, Statistical Genetics and Linear Models, Biomedical Machine Learning

New York University *Aug 2013-May 2015*
M.S. in Bioinformatics & Systems Biology
Overall GPA: 3.9/4
Thesis (supervisor: Christine Vogel): *iSUMO - integrative prediction of functionally relevant SUMOylated proteins*
Relevant courses: Linear Regression and Multivariate Analysis, Statistics in Biology, Bioinformatics and Genomes, Biological Databases and Data Mining, Proteomics Informatics, Mathematical Techiniques in CS Applications, Genomics and Global Health, Evolutionary Genetics and Genomics

Zhejiang University *Aug 2009-May 2013*
B.S. in Biotechnology
Overall GPA: 3.4/4
Thesis (supervisor: Ming Chen): *Optimizaiton of streptomycin production in S. avermitilis by metabolic network analysis*

EXPERIENCE

Weill Cornell Medicine Jul 2015 - Present
Graduate Research Assistant *New York, NY*

- Discovered three new patterns of complex structural variations in cancer genomes
- Designed, implemented, benchmarked *Junction Balance Analysis* to reconstruct genome graphs
- Conceptualized and developed *gGnome* for genome graph data structure
- Captured ongoing SV evolution in post-telomere crisis cell lines
- Characterized whole genome landscape in lung adenocarcinomas without RTK/RAS/RAF pathways
- Contributed to large cancer sequencing consortiums including The Cancer Genome Atlas (TCGA) and the Pan-Cancer Analysis of Whole Genomes (PCAWG)

New York University December 2013 - May 2015
Graduate Research Assistant *New York, NY*

- Trained predictive models for post-translational modification from protein function databases

New York University Feb 2014 - May 2015
Teaching Adjunct *New York, NY*

- Taught R statistical programming for Introduction to Biostatistics
- Tutored techniques for Biological Databases and Data Mining

3E Bioenergy Jun 2014 - May 2015
Bioinformatics Intern *New Brunswick, NJ*

- Compared crop genomes for candidate genes linked to drought resistance in sweet sorghum

RESEARCH PROJECTS

Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs May 2016 - Oct 2020
Senior leader: Imielinski M *New York, NY*

- Topology of junction copy number reveals novel classes of complex structural variants
- Rigma are deletion "chasms" at fragile sites arising early in GI tumor evolution
- Pyrgo are superenhancer-associated duplication "towers" in breast and ovarian cancer
- Tyfonas are "typhoons" of amplified fold-back inversions in acral melanoma

Illuminating rearranged cancer genome structure through gGnome Dec 2017 - Now
Senior leader: Imielinski M *New York, NY*

- Developed R API to genome graph data structures with a series of algorithms
- Systematically captured complex coding and non-coding SV driver events in pan-cancer genomes

Structural variant evolution after telomere crisis Dec 2019 - Sep 2020
Senior leader: de Lange T & Imielinski M *New York, NY*

- Screened more than a hundred shallow WGS for prevalent SV regions in clones of telomere crisis-surviving cells
- Reconstruct the exact linear allele after rearrangement with deep WGS
- Built consistent phylogeny using both SVs and SNVs
- Proved a single parental allele of chr12 to be the origin of SVs during crisis using allelic imbalances

Whole-genome characterization of lung adenocarcinomas lacking alterations in RTK / RAS / RAF pathways Dec 2017 - Sep 2020
Senior leader: Meyerson M, Govindan R, Imielinski M *New York, NY*

- Part of TCGA genomic data analysis network
- Discovered *KRAS* or *RTK/RAS/RAF* alterations (RPA) from WGS previously missed by whole exome sequencing
- Found higher *TP53* loss of function frequency in RPA- cancers
- Delinated diverse complex structural variation mechanisms creating amplification of oncogenes

PUBLICATIONS

Hadi K, **Yao X**, Behr JM, ..., Imielinski M. *Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs.* Cell. 2020;183: 197–210.e32.

Dewhurst SM, **Yao X**, ..., de Lange T, Imielinski M. *Structural variant evolution after telomere crisis.* BioRxiv. 2020. p. 2020.09.29.318436. doi:10.1101/2020.09.29.318436

(In review) Carrot-Zhang J, **Yao X**, Devarakonda S, et al. *Whole-genome characterization of lung adenocarcinomas lacking alterations in RTK/RAS/RAF/MAPK pathway.* Cancer Res. 2020;80: 5895–5895.

Wala JA, ..., Zhang C, Imielinski M, Beroukhim R. *SvABA: genome-wide detection of structural variants and indels by local assembly.* Genome Res. 2018. doi:10.1101/gr.221028.117

Gerstung M, Jolly C, Leshchiner I, ..., PCAWG Consortium. *The evolutionary history of 2,658 cancers.* Nature. 2020;578: 122–128.

Rheinbay E, Nielsen MM, ..., PCAWG Consortium. *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes.* Nature. 2020;578: 102–111.

Yao X, ..., Vogel C. *iSUMO - integrative prediction of functionally relevant SUMOylation events.* bioRxiv. 2017. p. 056564. doi:10.1101/056564

OPEN SOURCE SOFTWARE

Main author

- JaBbA - junction balance analysis
- gGnome - an R API to genome graphs

Contributor

- gGnome.js - an interactive web-based genome browser for genome graphs
- gUtils - elegant and fast genomic interval operations
- gTrack - static genome browser style plots
- fishHook - Gamma-Poisson regression for count data on genomic intervals
- GxG - interaction matrices between genomic bins

AWARDS

Asia Regional Winner, World 2nd Runner Up & Best New Application Nov 2011
2011 International Genetically Engineered Machines *Hong Kong, China; Cambridge, MA*

- Team member of ZJU-China
- Designed multicolor fluorescent expression system in biofilm responsive to gradients of oxygen level

Master's Student Research Grant 2014 & 2015
NYU Biology Master's Program *New York, NY*

- Funding for protien sumoylation prediction by mining public protein databases

Broad Institute Workshop Travel Grant 2015
Broad Institute *Cambridge, MA*

- Travel grant for single cell genomics workshop

SKILLS

Biology	Cancer Genomics, Computational Biology, Systems Biology
Sequencing Informatics	Illumina, Oxford Nanopore WGS
Statistics & Machine Learning	Generalized Linear Models, Random Forests, Regularized Regression
Computer Languages	R, Shell, Python, Java
Optimization	CPLEX, Gurobi
Data Visualization	ggplot, shiny
Scientific Communication	LaTeX, Adobe Illustrator, (R)markdown
Databases	SQLite, MySQL
Other Tools	Git, Emacs, Docker, Nextflow, Hugo